EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

(DOKTORA TEZİ)

VARLIK BAĞLAMA SİSTEMİNİN DİZİ ÖĞRENME YÖNTEMİ İLE BİLGİ TABANI KULLANILARAK GELİŞTİRİLMESİ

Emrah İNAN

Tez Danışmanı: Prof. Dr. Oğuz Dikenelli

Bilgisayar Mühendisliği Anabilim Dalı

Bilim Dalı Kodu: 619.01.00
Sunuş Tarihi: 04.Eylül.2018

Bornova-İZMİR
2018

Emrah İNAN tarafından **DOKTORA TEZİ** olarak sunulan "**VARLIK BAĞLAMA SİSTEMİNİN DİZİ ÖĞRENME YÖNTEMİ İLE BİLGİ TABANI KULLANILARAK GELİŞTİRİLMESİ**" başlıklı bu çalışma E.Ü. Lisansüstü Eğitim ve Öğretim Yönetmeliği ile E.Ü. Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve **04.09.2018** tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri Üyeleri:	**İmza**
Jüri Başkanı: Prof. Dr. Oğuz Dikenelli	
Raportör Üye: Doç. Dr. Rıza Cenk ERDUR	
Üye: Doç. Dr. Murat Osman ÜNALIR	
Üye: Doç. Dr. Belgin Ergenç BOSTANOĞLU	
Üye: Dr. Öğr. Üyesi Selma TEKİR	

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Doktora Tezi olarak sunduğum "VARLIK BAĞLAMA SİSTEMİNİN DİZİ ÖĞRENME YÖNTEMİ İLE BİLGİ TABANI KULLANILARAK GELİŞTİRİLMESİ" başlıklı bu tezin kendi çalışmam olduğunu, sunduğum tüm sonuç, doküman, bilgi ve belgeleri bizzat ve bu tez çalışması kapsamında elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara atıf yaptığımı ve bunları kaynaklar listesinde usulüne uygun olarak verdiğimi, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, bu tezin herhangi bir bölümünü bu üniversite veya diğer bir üniversitede başka bir tez çalışması içinde sunmadığımı, bu tezin planlanmasından yazımına kadar bütün safhalarda bilimsel etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim.

04/09/2018

Emrah İnan

ÖZET

VARLIK BAĞLAMA SİSTEMİNİN DİZİ ÖĞRENME YÖNTEMİ İLE BİLGİ TABANI KULLANILARAK GELİŞTİRİLMESİ

İNAN, Emrah

Doktora Tezi, Bilgisayar Mühendisliği Anabilim Dalı
Tez Danışmanı: Prof. Dr. Oğuz Dikenelli
04.09.2018, 87 sayfa

Son zamanlardaki Toplu Varlık Bağlama çalışmaları genellikle, anlamsal gömme ve çizge tabanlı yaklaşımlar kullanarak aynı metinde yer alan tüm eşlenmiş varlıkların küresel tutarlılığını sağlamaya çalışmaktadır. Çizge tabanlı yaklaşımlar başarılı sonuçlar gösterse de dayandıkları genel veri kümeleri için hesaplama açısından uzun süreli yaklaşımlardır. Ayrıca, anlamsal gömmeye dayanan çalışmalar dizileri düşünmeden sadece varlık çiftleri arasındaki ilişkiyi dikkate almaktadır. Bu tez kapsamında, bahsedilen problemlerin üstesinden gelebilmek için iki aşamalı bir yapay sinir modeli kullanarak ele alınmaktadır. İlk olarak, kolay söz-varlık çiftlerini eşleştirip, daha yakın atıfa sahip aday varlıkları filtrelemek için bu çiftin alan bilgisi kullanılmaktadır. İkinci aşamada iki taraflı Uzun Kısa Süreli Bellek ve küresel varlık anlamsızlığı için dikkat mekanizması kullanarak daha fazla anlam karmaşıklığı olan atıf-varlık çiftleri çözülmektedir. Bu tez kapsamında önerilen sistem oluşturulan alana özgü değerlendirme veri kümelerinde modern teknoloji sistemlerden daha iyi performans göstermektedir.

Anahtar Sözcükler: Varlık Bağlama, Dizi Öğrenme, Doküman Gömme, RDF Gömme, Bilgi Tabanı.

ABSTRACT

DEVELOPING ENTITY LINKING SYSTEM WITH SEQUENCE LEARNING BY USING KNOWLEDGE BASES

İNAN, Emrah

PhD. in Computer Engineering
Supervisor: Prof. Dr. Oğuz Dikenelli
04.09.2018, 87 pages

Recent collective Entity Linking studies usually promote global coherence of all the mapped entities in the same document by using semantic embeddings and graph-based approaches. Although graph-based approaches are shown to achieve remarkable results, they are computationally expensive for general datasets. Also, semantic embeddings only indicate relatedness between entity pairs without considering sequences. In this paper, we address these problems by introducing a two-fold neural model. First, we match easy mention-entity pairs and using the domain information of this pair to filter candidate entities of closer mentions. Second, we resolve more ambiguous pairs using bidirectional Long Short-Term Memory and attention mechanism for the global entity disambiguation. Our proposed system outperforms state-of-the-art systems on the generated domain-specific evaluation dataset.

Keywords: Entity Linking, Sequence Learning, Document Embeddings, RDF Embeddings, Knowledge Base

TEŞEKKÜR

Öncelikle doktora sürecinde desteğini hiçbir zaman esirgemeyen ve vizyonunu aşılayan danışmanım Prof. Dr. Oğuz DİKENELLİ başta olmak üzere; tez çalışmasının şekillenmesinde değerli görüş ve önerileri ile katkıda bulunan jüri üyesi hocalarım Doç. Dr. Murat Osman ÜNALIR ve Doç. Dr. Belgin Ergenç BOSTANOĞLU'na çok teşekkür ederim.

Bölümümüz Seagent Laboratuvarındaki değerli çalışma arkadaşlarımla birlikte Vahab MOSTAFAPOUR ve FATİH TEKBACAK'a doktora sürecinde sağladıkları teknik ve manevi destek ile beni yalnız bırakmadıkları için teşekkür ederim.

Hayatım boyunca karşılaştığım her zorlukta beni koşulsuzca destekleyen, bugünlere gelmemi borçlu olduğum aileme çok teşekkür ederim.

Son olarak, 2211-Yurt İçi Doktora Burs Programı kapsamında beni destekleyen Türkiye Bilimsel ve Teknik Araştırma Kurumu'na (TÜBİTAK) teşekkür ederim.

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (Devam)

İÇİNDEKİLER (Devam)

İÇİNDEKİLER (Devam)

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltma	**Açılım**
BT	Knowledge Base *(Bilgi Tabanı)*
DOC2VEC	Document to Vector Embeddings *(Dokümandan Vektör Gömme)*
KBP	Knowledge Base Population *(Bilgi Tabanı Üretimi)*
LSA	Latent Semantic Analysis *(Gizil Anlamsal Analiz)*
UKHA	Long Short Term Memory *(Uzun Kısa Terim Hafızası)*
NER	Named Entity Recognition *(Varlık Tanımlama)*
NIST	National Institute of Standards and Technology *(Ulusal Standartlar ve Teknoloji Enstitüsü)*
DBLP	Digital Bibliography and Library Project *(Dijital Kaynakça ve Kütüphane Projesi)*
OWL	Web Ontology Language *(Veb Ontoloji Dili)*
POS	Part Of Speech *(Konuşmanın Parçası)*
RDF	Resource Description Framework *(Kaynak Tanımlama Çerçevesi)*
RDF2VEC	RDF to Vector Embeddings *(RDF'ten Vektör Gömme)*
SPARQL	SPARQL Protocol and RDF Query Language *(SPARQL Protokolü ve RDF Sorgu Dili)*
TAC	Text Analysis Conference *(Metin Analizi Konferansı)*
URI	Uniform Resource Identifier *(Birleşik Kaynak Tanımlayıcısı)*
VB	Entity Linking *(Varlık Bağlama)*
YAGO	Yet Another Great Ontology *(Bir Diğer Büyük Ontoloji)*
WSD	Word Sense Disambiguation *(Kelime Anlamı Çözümleme)*

1 GİRİŞ

Bu bölümde tezin motivasyonu ve problem tanımı yapılarak tezin kapsamındaki araştırma alanı incelenmiştir. Daha sonra araştırma alanı olan yöntem ve önerilen çözüm tartışılarak tezin literatüre katkıları açıklanmıştır. Son olarak, tezin yapısı ileriki bölümlere genel bir bakış ile ortaya konulmuştur.

1.1 Motivasyon ve Problem Tanımı

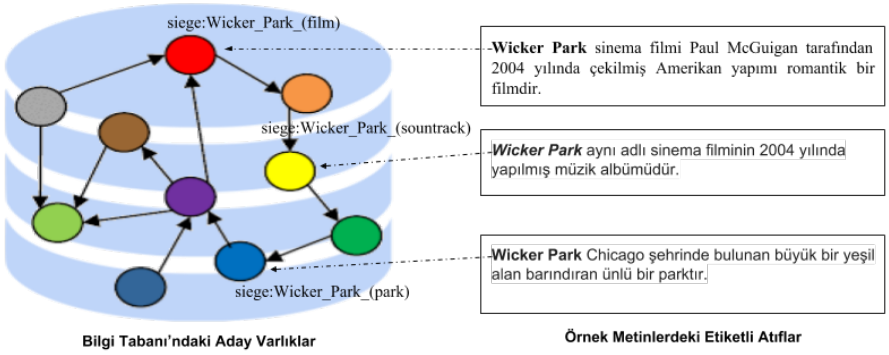
Son yıllarda internet çok hızlı büyüyerek bir çoğumuz için en önemli bilgi kaynağı olmuştur. Mevcut haliyle internet 20 milyardan fazla veb sayfası[1] barındırmaktadır. Ancak bu bilgi kaynağının büyük bir kısmı yapısal olmayan bir şekilde bulunmaktadır. İnternet verisinin yapısal bir biçime çevrilmesi ve veri kaynaklarını birbiri ile anlamsal olarak bağlanması "Veri Vebi" araştırma alanını oluşturmaktadır (Shadbolt et al., 2006). Verilerin yapısallaştırılması ile sezgisel analiz (Melville et al., 2009), soru cevaplama (West et al., 2014) ve sohbet robotu (Huang et al., 2007) sistemlerinin daha etkin bir şekilde çalışmasına olanak tanımaktadır. Yapısallaştırılan veriler birbiri ile iyi belirlenmiş hiyerarşi ve diğer anlamsal ilişkilerle bağlanmış gerçek dünya varlıklarının oluşturduğu bilgi tabanlarını ortaya çıkarmıştır.

Bağlı Veri uzayı içinde yer alan bilgi tabanı kaynaklarından Vikipedi[2] yarı yapısal bir bilgi kaynağı olarak işbirlikli kullanıcılar tarafından devamlı olarak güncellenmektedir. Vikipedi kaynağının yapısallaştırılması ile elde edilen DBpedia (Bizer et al., 2009) Bağlı Veri uzayının merkezinde yer almaktadır. Güncel haliyle Bağlı Veri uzayı DBpedia ile bağlantılı toplam 1184 bilgi tabanı barındırmaktadır[3]. En ünlülerinden BabelNet (Navigli and Ponzetto, 2012a) ve YAGO (Suchanek et al., 2007) DBpedia'nın farklı yöntemlerle zenginleştirilmesiyle oluşturulmuştur.

İnternet üzerinde bulunan yapısal olmayan metinlerin yapısal bilgi tabanları ile bağlanması Anlamsal Veb vizyonu açısından kritik bir hedeftir. Bu hedefe ulaşmanın kritik ve başlangıç adımı Varlık Bağlama yöntemleridir. Varlık Bağlama probleminin çözümü ilişki çıkarımı (Weston et al., 2013), bağ tahmini (Nickel et al., 2015) ve bilgi tabanı tamamlama (Minervini et al., 2016) gibi araştırma alanları için öncü bir adımdır. Varlık Bağlama, metindeki atıfların seçilen bilgi tabanındaki ilgili tanımlı varlıklarla etiketlenmesidir. Burada atıf özel ismin verilen metindeki

[1] http://www.worldwidewebsize.com/
[2] https://tr.wikipedia.org/wiki/Vikipedi
[3] http://lod-cloud.net/about

geçtiği halini gösterirken tanımlı varlık atıfın bilgi tabanındaki referans bağlantısını belirtmektedir.



Şekil 1.1: Varlık Bağlama probleminin tanımı

Varlık bağlama probleminin ve kavramlarının daha detaylı bir biçimde açıklanması için Şekil 1.1 örnek gösterilmiştir. ilk cümlede yer alan **Wicker Park** atıfı verilen Bilgi Tabanı'ndaki **siege:Wicker_Park_(film)** varlığı ile bağlanmıştır. Burada **Wicker Park** atıfı üç ayrı cümlede geçmiştir ve üç farklı aday varlığa sahiptir. Atıfın her cümlenin bağlamına göre verilen bilgi tabanındaki ilgili varlıkla eşlenmesi Varlık Bağlama'nın temel problemidir. İkinci cümlede aynı isimle yer alan atıfın **siege:Wicker_Park_(soundtrack)** aday varlığı ile eşlenmişken üçüncü cümlede **siege:Wicker_Park_(park)** varlığı ile bağlanmıştır.

Varlık Bağlama yöntemleri, atıfların etiketlenmesi, aday varlıkların oluşturulması (AVO) ve aday varlık sıralanması (AVS) olmak üzere üç aşamadan oluşmaktadır. İlk adım atıfların etiketlenmesi için metinde geçen atıfların tespit edilmesi aşamasıdır. Atıfların tespiti için Stanford CoreNLP (Manning et al., 2014) ve benzeri araçlar ile metinlerde yer alan kelimelerin tipi (isim, fiil, sıfat, vb.) belirlenmektedir. Belirlenen kelime ve kelime grupları Bilgi Tabanı'nda sorgulanarak aday varlıklar oluşturulmaktadır. Bir atıfı işaret eden birden fazla aday varlık olabileceği durumu göz önünde bulundurularak en uyumlu atıf-varlık eşleşmesi aday varlıkların sıralanması aşamasında yapılmaktadır. Son adıma aday varlıkların sıralanması ile birlikte varlık çözümleme de denmektedir.

Atıfların etiketlenmesi ve aday varlıkların sıralanması aşamaları Varlık Bağlama yöntemlerinin temel amaçlarıdır. Atıfların etiketlenmesi aşamasının temel problemi, herhangi bir doğal dil işleme ve benzeri aracı olmayan doğal dillerde daha çok önem kazanmaktadır. Bunun için genel olarak metinlerin kelime listeleri çıkarılarak Bilgi Tabanı'nda sorgulanması çözümü ön plana çıkmaktadır. Aday varlıkların sıralanması aşamasının temel problemi ise var olan atıf için aday varlıklar arasından en uygun varlığın bağlanmasıdır. Şekil 1.1

üzerinde gösterilen üç cümlede Wicker Park atıfına ait verilen bilgi tabanında **siege:Wicker_Park_(film)** varlığına ek olarak filmin müziklerini içeren albüm olan **siege:Wicker_Park_(soundtrack)** ve Chicago eyaletinde yer alan halka açık bir park **siege:Wicker_Park_(Chicago_park)** aday varlıklarına sahiptir. Bu aşamanın temel amacı örnek cümlelerin metinsel bağlamına göre doğru atıf-varlık eşleşmesinin yapılmasıdır. Böylece, ilk cümlenin bağlamı sinema alanında olmasından dolayı film olan varlıkla eşleşmesini sağlamaktır.

Aday varlıkların sıralanması için varlıklar arasındaki anlamsal ilintilik hesabı ön plana çıkmaktadır. Anlamsal ilintililik tanımlanırken benzerlik ile birlikte ele alınması daha uygun olmaktadır. İlintililik birbiriyle ilintili olan ya da ortak bir ilişkiye sahip olan anlamına gelmektedir (Cochez et al., 2017). Diğer yandan benzerlik ise ortak karakteristiğe sahip olmayı gösteren ilintililiğin özel bir durumudur. Örnek olarak "Google" ve "Sergey Brin" kurucusu olma ilişkisi üzerinden birbiriyle ilintili iki varlıktır. Ancak kavramsal olarak organizasyon ve insan oldukları için benzer varlıklar değildir.

Bilgi tabanı kullanıldığı durumda anlamsal ilintililik iki varlığın birbirlerine olan anlamsal yakınlıklarının hesaplanması olarak tanımlanabilir. Anlamsal yakınlık hesabı için iki varlık arasındaki düğüm ve ilişki sayısı ile izlenen patikanın hesaplandığı çizge tabanlı yaklaşımlar vardır (Schuhmacher and Ponzetto, 2014). En çok kullanılan ilintililik hesaplarından biri Vikipedi bağ yapısı sayesinde iki aday varlığa gelen ve kendilerinden giden bağlarla yapılan oranlamadır (Milne and Witten, 2008). Aday varlıkların sıralanmasında Varlık Bağlama sistemleri her bir atıfı aday varlıkları arasında en yüksek anlamsal ilintililiğe sahip varlığı seçerek eşleştirmektedir. Aday varlıkların anlamsal ilintililik değerlerine göre sıralanması çözümleme aşamasındaki en kritik adımdır. Bu ilintiliği hesaplamak için güncel yöntemlerden TAGME (Ferragina and Scaiella, 2010) ve Cucerzan (Cucerzan, 2007a) gibi alan bağımsız Varlık Bağlama çalışmalarının çoğunluğu Vikipedi bağ yapısına dayanan anlamsal ilintililik yöntemini kullanmaktadır. Alan bağımlı bilgi tabanı geliştirildiğinde ise bu bağ yapısı kapsamlı bir bilgi kaynağı sunmaktan uzaklaşmakta ve sistemin genel başarısını azaltmaktadır. Bu yüzden, bilgi tabanından bağımsız yöntemlerin gerekliliği alana özgü ortamda ortaya çıkmaktadır.

Vikipedi bilgi kaynağı Varlık Bağlama için önemli olmasına rağmen bilgi tabanından bağımsız yaklaşımların global varlık çözümleme aşamasında kullanılması önemli bir gereksinimdir. İstenilen bilgi tabanı DBpedia (Mendes et al., 2011) örneğinde görüldüğü üzere anlamsal yapısı güçlü olabileceği gibi BabelNet (Navigli and Ponzetto, 2012b) kaynağındaki gibi çizge temelli de

olabilir. Gereksinim alan bağımlı bir bilgi tabanına ihtiyaç duyarsa Varlık Bağlama sisteminin DBLP (Ley, 2002) ve LinkedMDB (Hassanzadeh and Consens, 2009) bilgi kaynakları üzerinde de çalışması beklenmektedir.

Kelime gömme (Mikolov et al., 2013a) yöntemlerinin ortaya çıkmasıyla bilgi tabanı bağımsız Varlık Bağlama yöntemlerinin önü açılmıştır. Kelime gömme modelleri daha genelleştirilmiş tutarlılık ve ilintililik hesapları yapabildiği için herhangi bir bilgi tabanına ve elle yapılmış kurallara ihtiyaç olmadan çalışabilmektedir. Böylece bilgi çok bilinen tabanının yapısına ihtiyaç duymadan herhangi bir yapıda çalışabilecek sistemler geliştirilmiştir (Zwicklbauer et al., 2016a; Usbeck et al., 2014).

En son sinir ağı modelleri atıf ve varlıklar için benzerlik ve ilintilik hesaplarında daha genelleştirilmiş yaklaşımlar kullanmıştır (Sun et al., 2015). Aynı zamanda atıf ve varlıklarını vektörel temsillerini aynı uzayda birleştiren çalışmalar vardır (Yamada et al., 2016). Bu tür çalışmalar bütün gömme modellerini tek bir vektör uzayına koyduğu için hesaplama maliyetleri yüksek olmaktadır. Diziden diziye öğrenme modelleri (Sutskever et al., 2014) ortaya çıkması ve sinir ağı makine çevirisinde gösterdikleri başarıdan dolayı global varlık çözümlemenin dikey katmanlar olarak çözülmesinin önü açılmıştır (Bahdanau et al., 2014). Orjinal sinir ağı makine çeviri modelinde koşullu olasılık kullanılarak kaynak cümlenin başka bir doğal dildeki karşılığı olan hedef cümle tahminlenmektedir. Benzer şekilde atıflar dizisi dikey katmanlar sayesinde hedef varlık dizisinin tahminlenmesi global varlık çözümleme yöntemi olarak geliştirilebilinir.

Alan bilgisi belirli bir iş alanına ait gerçeklerin tamamı olarak tanınlanabilir. Bu yüzden alan uyumlandırılması Varlık Bağlama sistemleri için kritik bir görev olarak verimli bir şekilde hesaplama maliyetini düşürebilmektedir. İşlenen verinin verimli bir şekilde düşmesi global çözümleme yönteminin daha etkili çalışmasına olanak sunmaktadır. Buna ek olarak alan uyumlandırılması bilgi yoğun alanlar için performansı artıran bir etkendir (Navigli, 2013). AIDA-light (Nguyen et al., 2014a) öncül bir çalışma olarak alan bilgisini belirli bir bilgi tabanı için varlık çözümleme aşamasında kullanmıştır. Ancak bu çalışma bilgi tabanı bağımsız gömme modellerine dayanmamaktadır.

Bu tez kapsamında önerilen GnoSSEA adıyla anılan Varlık Bağlama yöntemi diziden diziye öğrenme modelini global varlık çözümleme aşamasına uyarlamaktadır. Bilgi tabanında ve belgelerden bağımsız olarak anlamsal gömme (Ristoski and Paulheim, 2016) modeli çift yönlü Uzun Kısa Terim (Bi-UKHA) (Graves and Schmidhuber, 2005) algoritmasının girdisi olarak kullanılmaktadır.

Bu sebeple GnoSSEA atıf ve varlıkların vektörel temsilleri üzerinde çalıştığı için bilgi tabanı bağımsız bir yöntem sunmaktadır. Aynı zamanda dikkat mekanizması (Bahdanau et al., 2014) alan bilgisine göre öne çıkan aday varlıkların ağırlıklarını artırarak çözümleme aşamasını daha verimli hale getirmektedir.

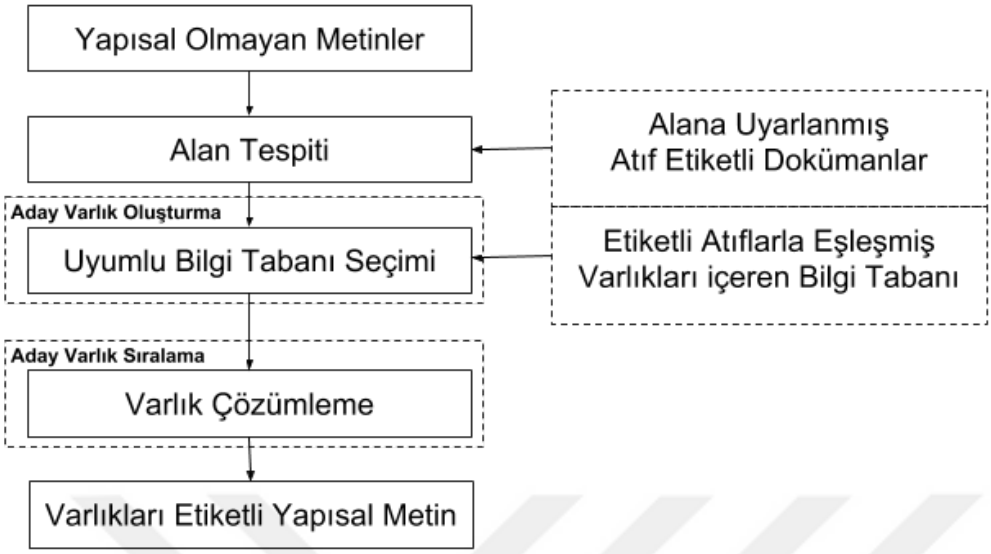
Tezin bütününe bakıldığında özgün yaklaşımlardan alan uyumluluk aşaması diziden diziye öğrenme modelinden önce metne ait alan bilgisinin katman olarak değerlendirilmesidir. Böylece yeni bir alanın eklenmesi ile genişleyen ölçeklenebilir bir mimari sunulmaktadır. Veri boyutunu düşürmek ve daha az ama daha etkin verileri girdi olarak kullanmak için iki adımlı bir yöntem geliştirilmiştir. İlk önce "kolay etiket" adı altında atıf-varlık çiftlerinin eşleştirilmesi yapılmıştır. Bu adımda sadece bir aday varlığı olan atıfların eşleştirilmesi ile hem veri sayısı azaltılıp hem de metnin alan bilgisi için bir ipucu yakalanmaktadır. İkinci aşamada metnin alan bilgisi bu ipucu ve belge gömme modeliyle tespit edilerek diziden diziye öğrenme modeli ile global varlık çözümleme gerçekleştirilmiştir.

1.2 Tezin Genel Çerçevesi

Bu tez kapsamındaki çalışma, Varlık Bağlama probleminin temel hedefi olan aday varlıkların sıralanması için sinir ağlarının daha gelişmiş bir hali dizi öğrenme modeli ile bir çözüm önermektedir. Atıf tespiti için son zamanlardaki çalışmalara benzer şekilde (Zwicklbauer et al., 2016a; Usbeck et al., 2014), önceden tespit edilmiş atıflar içeren metinlerin bu tez kapsamında önerilen yöntemin girdileri olduğunu varsayıyoruz. Ayrıca, aday varlıkların oluşturulması aşamasında her bir atıfa ait verilen bilgi tabanında bir veya daha fazla aday varlıkların olduğu varsayılmaktadır.

Şekil 1.2 tez kapsamında sunulan yöntemin genel üst yapısını göstermektedir. Yapısal olmayan metinler için alanların tespiti aşamasında Vikipedi bilgi kaynağındaki alana uyarlanmış atıf etiketli metinleri alarak doküman gömme modeli eğitilmektedir. Vikipedi kategorisi sayfaları ilgili belgeleri çıkarmak için kullanılmıştır. Daha sonra bu belgeler ile Doc2Vec (Le and Mikolov, 2014) modeli ile eğitilmektedir. Verilen metnin alanı tespit edildikten sonra, dizi öğrenme yöntemleri kullanılarak varlıklar çözümlenmektedir. Bu çalışmada DBpedia bilgi tabanı RDF gömme modelinin girdilerini kolayca sağlamak amacıyla alana özgü bilgi tabanlarını çıkarmak için kullanılmaktadır.

Bu tezin kapsamında önerilen Varlık Bağlama yöntemi diziden diziye öğrenme modeline alan bilgisini harmanlayan yeni bir mimari yapısı içermektedir.



Şekil 1.2: Varlık Bağlama yönteminin genel mimarisi

Bu mimari dizi öğrenme modellerinden iki yönlü UKHA (Graves and Schmidhuber, 2005) ve dikkat mekanizmasını (Bahdanau et al., 2014) kullanmaktadır. Gerçekleştirilen bu modeller için alan uyumlu anlamsal ve doküman gömme modülleri oluşturulmuştur. Bu modülleri oluşturmak adına önceden eşlenmiş atıf-varlık çiftlerini barındıran atıf etiketli belge yığını olarak Vikipedi kaynağı kullanılırken bu atıflara eşleşmiş varlıkları içeren DBpedia bilgi tabanı sinema, kitap ve müzik olmak üzere seçilen üç alan için ayrı ayrı çıkarılmaktadır.

Doküman ve anlamsal gömme modülleri sırasıyla alan tespiti ve global varlık çözümleme aşamalarında kullanılmaktadır. Alan tespiti için kullanılan yöntem Doc2Vec (Le and Mikolov, 2014) modeli olmuştur. Birçok alan tespiti yöntemi olmasına rağmen güncel olan Doc2Vec modeli diğer yöntemlere karşı başarı sağladığı için bu tez kapsamında bu model üzerinde durulmuştur. Yapısal olmayan metin bu aşamadan geçtiğinde bağlam olarak hangi alanda olduğu tespit edilmesi ile birlikte bir sonraki aşama olan global varlık çözümleme aşamasına geçilmektedir. Global varlık çözümleme aşamasında alan tespiti yapıldığı için aday varlıkların sayısı azalmış olmaktadır. Bu durumda global varlık çözümleme aşamasını belirli bir alan için yapılmasından dolayı işlem süresini azaltmaktadır.

Tezin son aşamasında önerilen yöntem güncel ve veb servisi olan Varlık Bağlama sistemleri ile popüler bir çevrimiçi değerlendirme aracında karşılaştırılmıştır. Karşılaştırmanın alan uyumlu veri kümelerinde yapılabilmesi için geliştirilen bir yöntem ile seçilen üç alanı da kapsayan anlam karmaşıklığı düşük ve yüksek olmak üzere iki farklı değerlendirme kümesi hazırlanmıştır (Inan and

Dikenelli, 2017). Vikipedi kaynağındaki anlamsal kategori sayfalarındaki bilgilerle DBpedia kaynağında yer alan şema bilgileri örtüştürerek alan bilgilerine göre iki kaynak için ayrı ayrı alan uyumlu veri kaynakları üretilmiştir.

1.3 Tezin Katkıları

Bu tezin ana amacı alana uyarlanmış bilgi tabanlarını ve etiketli metin depolarını kullanarak diziden diziye öğrenme modelini ve alan bilgisini harmanlayan bir Varlık Bağlama metodu sunmaktır. Tez kapsamında yapılan katkıları alana uyumlu Varlık Bağlama yöntemi ile birlikte öğrenme ve değerlendirme veri kümelerinin oluşturulması açısından iki ana başlık altında toplanmıştır.

- Alana uyumlu Varlık Bağlama yönteminin tez kapsamındaki katkıları
  - Sunulan Varlık Bağlama yönteminde yer alan anlamsal ve doküman gömme modellerini üretmek için girdi olarak herhangi bilgi tabanı ve belge yığınından yararlanmaktadır.
  - Bu yöntem herhangi bir bilgi tabanı ya da doküman yığını kullanılarak alan tespiti ve alan bilgisi harmanlanmış global varlık çözümleme aşamalarında kullanılma olanağı sunmaktadır.
  - Sunulan diziden diziye öğrenme modeli üç katmanlı bir mimari gerçekleştirilmiştir. Bu mimari Yapay Sinir Ağları Çeviri probleminin herhangi bir dile ait cümleleri yerine atıf dizilerini kaynak dizi olarak ele almaktadır. Ele alınan bu diziden referans varlıkların dizisi tahminlenmektedir.
  - Tez kapsamında sunulan özgün alan bilgisi katmanı yapay sinir modellerinden önce kullanıldığından Varlık Bağlama için genişletilebilir bir mimari önermektedir. Böylece gerekli durumlarda yeni bir alana ait bilgi tabanı ve doküman derlemi eklenerek genişletilebilmektedir.
  - Son olarak geliştirilen yönteme adapte edilen dikkat mekanizması alan bilgisini kullanarak aday varlıklardan alana özgü olanlarının ağırlıklarını artırarak atıf-varlık eşleme açısından ağırlığını yükseltmektedir.
- Alana özgü veri kümeleri oluşturma açısından yapılan katkılar
  - Alana özgü veri kümeleri oluşturmak için bu tez kapsamında bir araç geliştirilmiştir. Verilen doküman yığınındaki kategorik

bilgiler doğrultusunda hem öğrenme hem de değerlendirme kümeleri oluşturulmuştur.

- Kategorilerine göre ayarlanmış bu dokümanlar içindeki atıflarla eşleşen varlıkların barındığı bilgi tabanları üzerinden doküman gömme ve anlamsal gömme modelleri elde edilmiştir.
- Değerlendirme kümesi olarak oluşturulan bu veri kümeleri çok bilinen Varlık Bağlama değerlendirme araçlarına uyumlu olarak geliştirildiği için kolaylıkla bu araçlar üzerinde başka Varlık Bağlama araçlarının başarısı gözlenebilmektedir.
- Son olarak geliştirilen bu araca başka doğal dillere ait bilgi tabanı ve doküman yığınlarının eklenmesi ile herhangi bir doğal dil için de aracın kullanılması sağlanmaktadır.

1.4 Tezin İçeriği

Tezin ikinci bölümünde Varlık Bağlama probleminin temelleri için genel bir bakış açısı sunulmuştur. Bu problemin başlangıcında hangi adımların çözümlenmesi gerektiği ve geliştirilen yöntemin değerlendirilmesi için gereken veri kümeleri ve sonuçların karşılaştırılacağı ölçütleri detaylandırılmaktadır.

Tezin üçüncü bölümünde Varlık Bağlama sistemlerinin tarihsel süreçte izlediği yolu göz önünde bulundurarak literatür taraması olarak sunulmaktadır. Varlık Bağlama yöntemlerinin genel yapıları karşılaştırılmış ve tezde sunulan yöntemin ayrıştığı yerler netleştirilmektedir.

Tezin dördüncü bölümünde diziden diziye öğrenmeye dayanan alana özgü bir Varlık Bağlama yöntemi sunulmuştur. Aynı zamanda alan bilgisi kullanılarak aday varlıkların sayısı filtrelenmiş ve bir tane aday varlığı olan kolay atıf-varlık eşleşmeleri de çözümleme aşamasında bağlamı tespit etme aşamasında kullanılmaktadır.

Beşinci bölümde tez kapsamında önerilen yeni yöntemin son teknoloji Varlık Bağlama sistemleri ile karşılaştırılması yapılmaktadır. Bunun için değerlendirme veri kümelerinin alana özgü olarak oluşturulmasından değerlendirme sonuçları karşılaştırılmaktadır.

Son bölümde tezin özeti yapılarak gelecek çalışmalar hakkında planlar aktarılmaktadır. Tez kapsamında yapılan önemli katkılar ile birlikte yapılan

kabullenme ve kısıtlamalar detaylandırılmış ileriki çalışmalar için geliştirilebilecek yönleri özetlenmektedir.

1.5 Yayınlar

Bu tez kapsamında konferans ve çalıştay yayınları yapılmıştır. Ayrıca hakem yorumlarının beklenildiği bir dergi makalesi de bulunmaktadır. Bu yayınlar tezin kapsamı çerçevesinde aşağıdaki şekilde sıralanmıştır.

- **GnoSSEA: A Sequence-to-Sequence Domain Oriented Entity Linking System (Under Review) (Alana Özel Varlık Bağlama için Dizi Öğrenme Metodu (İncelemede)** Bu çalışma tezde önerilen ana yöntemi barındıran diziden diziye öğrenme modellerinin genişletilerek denendiği çalışmadır. Burada sinema alanına ek olarak kitap ve müzik alanları da dahil edilerek alana uygun ölçeklenebilir bir mimari sunulmaktadır. Yeni bir alana özgü belge ve anlamsal gömme modellerinin eklenebileceği bu yöntem UKHA ve Bi-UKHA sinir ağı modellerinde eğitilmiştir. Daha sonra alan bilgisinin eklendiği bir dikkat mekanizmasından geçerek anlam karmaşıklığı yüksek ve düşük olarak iki farklı şekilde ayarlanmış üç alandan harmanlanan değerlendirme kümelerinde son teknoloji sistemlerle karşılaştırılmıştır.

  Emrah Inan, Oguz Dikenelli: GnoSSEA: A Sequence-to-Sequence Domain Oriented Entity Linking System. Knowledge-based Systems Journal (Under Review)

- **A Sequence Learning Method for Domain-Specific Entity Linking (Alana Özel Varlık Bağlama için Dizi Öğrenme Metodu)** Bu çalışma tezin ana yönteminin bir parçasını barındıran Dizi Öğrenme modellerinin ilk denendiği çalışmadır. Burada sinema alanına özgü üretilmiş bilgi tabanında anlamsal gömme ile elde edilmiş girdiler UKHA ve Bi-UKHA sinir ağı modellerinde eğitilmiştir. Daha sonra sinema alanında anlam karmaşıklığı yüksek olan değerlendirme kümesinde güncel yöntemlerle karşılaştırılmıştır.

  Emrah Inan, Oguz Dikenelli: A Sequence Learning Method for Domain-Specific Entity Linking. NEWS@ACL 2018: 14-21

- **WeDGeM: A Domain-Specific Evaluation Dataset Generator for Multilingual Entity Linking Systems (Çok dilli Varlık Bağlama sistemleri için alana özgü değerlendirme veriseti üretim aracı)** WeDGeM aracına

Vikipedi İngilizce yığınları[4] otomatik olarak konsoldan indirilmiştir. Daha sonra bu yığınlar birden çok iş parçacığı yapısına sahip Akka[5] aracı ile SAXParser sayesinde işlenmiştir. İşlenen veri Vikipedi URI ve etiketli metin olarak doküman veri deposuna yüklenmiştir. Daha sonra verilen bilgi tabanındaki varlıklara uygun alandaki metinlerle eşlenerek mevcut atıf ve varlıklara sahip etiketli metinler elde edilmiştir. Bu tez kapsamında Varlık Bağlama yöntemi iki ana yapıdan oluşmaktadır. İlki alan bağımlı bilgi tabanından üretilen anlamsal gömmelerin varlık çözümlemede kullanılmasıdır. İkinci yapı da alan bağımlı Varlık Bağlama yaklaşımlarını değerlendirecek veri kümeleri bulunmadığı için bu tür veri kümelerinin üretildiği bir sistemin geliştirilmesinden oluşmaktadır.

> Emrah Inan, Oguz Dikenelli: WeDGeM: A Domain-Specific Evaluation Dataset Generator for Multilingual Entity Linking Systems. WISE (2) 2017: 221-228

- **Effect of Enriched Ontology Structures on RDF Embedding-Based Entity Linking (RDF gömme tabanlı Varlık Bağlama'da zenginleştirilmiş ontoloji yapısının etkisi)** Bu çalışma kapsamında DoSeR çalışmasının geleneksel yöntemlerin performanslarından daha yüksek başarı göstermesinden dolayı ontoloji yapılarında hangi bileşenlerin zenginleştirilmesi genel performansa etki edildiği gözlenmiştir. Anlamsal (RDF) izdüşümleri son zamanlarda Varlık Bağlama sistemlerinde aday varlıklar arasından atıf için en uygun tanımlı varlığı çözümlemek için kullanılmaktadır. Bu bölümde varlıkların çözümlenmesi aşamasında RDF izdüşümlerine dayalı model ile anlamsal ilintililik hesaplanmıştır. Daha sonra hangi zenginleştirilmiş ontoloji yapısının RDF izdüşümlerine dayanan Varlık Bağlama yaklaşımlarındaki başarıyı daha çok artırdığı gözlenmiştir. Alan bağımlı bilgi tabanı HDT[6] aracı sayesinde hızlı bir şekilde sorgulanarak atıf ve varlıklara ait URI bilgileri doğrulanarak indeks oluşturulmuştur. Oluşturulan bu indeks etiketi yani atıfı anahtar olarak saklanırken bilgi tabanından çekilen URI bilgisi de değer olarak anahtar-değer veri deposunda saklanmıştır. WeDGeM aracından üretilen etiketli metinlerle uyumlu bu indeks daha sonra aday varlıkların üretilmesi aşamasında kullanılmıştır.

> Inan E., Dikenelli O. (2017) Effect of Enriched Ontology Structures on RDF Embedding-Based Entity Linking. In: Garoufallou E., Virkus S., Siatri R., Koutsomiha D. (eds)

[4] https://dumps.wikimedia.org/en/20171103/
[5] https://akka.io/
[6] http://www.rdfhdt.org/

Metadata and Semantic Research. MTSR 2017. Communications in Computer and Information Science, vol 755. Springer, Cham

Aday varlıklar elde edilmesine paralel olarak bilgi tabanı varlık-ilişki dizilerine çevrilmiştir. Bu diziler daha sonra anlamsal izdüşüm hesaplamak için RDF2Vec modelinde eğitilmiştir. Burada Word2Vec (Mikolov et al., 2013b) modeli RDF izdüşümlerine dönüştürerek varlıklar arası anlamsal ilintiliği hesaplamaktadır. Varlıkların çözümlenmesi için AGDISTIS (Usbeck et al., 2014) çalışmasında kullanılan HITS algoritmasına ek olarak DoSeR (Zwicklbauer et al., 2016a) çalışmasındaki PageRank algoritması en popüler yöntemlerdendir. Burada Doc2Vec modeline dayanan özgün bir yöntem de denenecektir. Sonuç olarak atıflara ait varlıkların URI bilgilerinin tutulduğu JSON biçiminde etiketli metinler elde edilmiştir. Örnek olarak metinde geçen "Wicker Park" atıfı park yada albüm olan diğer aday varlıkların çözümlenmesinden sonra film olan Wicker Park varlığına ait URI bilgisi ile eşleşmiştir.

2 VARLIK BAĞLAMA TEMELLERİ

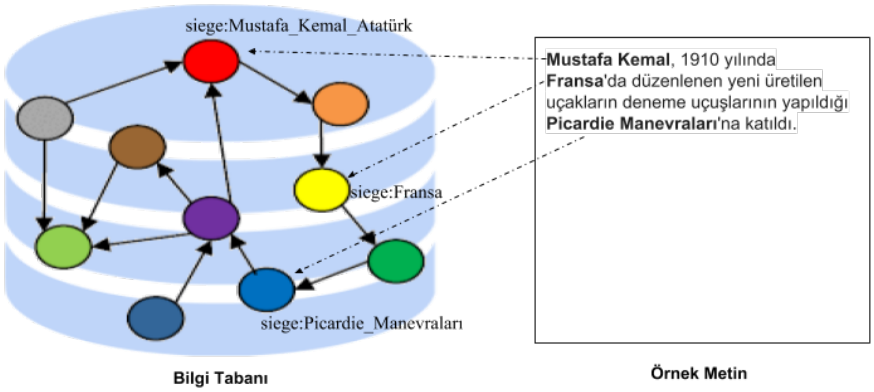
Yapısal olmayan metinlerin yapısallaştırılması için temel adım olan Varlık Bağlama problemi son on yılda kapsamlı bir şekilde incelenmiştir (Shen et al., 2015). Varlık Bağlama internet üzerindeki metinlerin yapısallaştırma adımı olarak düşünüldüğünde bir çok araştırma alanının da temel adımı olarak görülmektedir. Örnek olarak varlıklar arasındaki ilişkileri bulmayı amaçlayan İlişki Çıkarımı (Weston et al., 2013) varlıkları etiketlenmiş metinlere ihtiyaç duymaktadır. Ayrıca Bağ Tahminleme (Nickel et al., 2015) ve Bilgi Çizgesi Tamamlama (Minervini et al., 2016) araştırma konuları için de varlıkların önceden bağlanması gerekmektedir. Bu bölümde, Varlık Bağlama ve onun temelleri daha spesifik olarak, farklı problemleri içeren genel bir bakış açısı sunulmaktadır.

Bilgi çıkarımı[7] tekniklerinin gelişmesi ve Vikipedi gibi veb tabanlı ansiklopedilerinin oluşması geniş ölçekli Bilgi Tabanı kaynaklarının kurulmasının önünü açmıştır. Bir metnin anlamını otomatik olarak anlamak, DBpedia (Auer et al., 2007) ve Freebase (Bollacker et al., 2008) gibi Bilgi Tabanı kaynaklarının ortaya çıkmasıyla birlikte önem kazanmıştır. Veri Veb'i açısından irdelendiğinde yapısal olmayan veb dokümanları yapısal Bilgi Tabanı kaynaklarına etiketlenerek köprü kurulmasıyla daha anlamlı hale getirilmekte bu da Bağlı Açık Veri bulutunun evrimleşmesine yardımcı olmaktadır. Tanımlı Varlıklar insan, organizasyon veya yer isimleri gibi elle tutulur, gerçek ve özel isim atanan kavramlardan oluşurken Kelime Çözümleme de tek anlamlı ya da çok anlamlı bir veya birden fazla kelime grubu kümesini içermektedir.

Varlık Tanımlama (VT), yeni terimleri tanımlarken ontolojik tipi (insan, organizasyon ve yer) olarak sınıflandırmaktadır. Ayrıca tanımladığı terimlerin girilen mevcut metin üzerindeki başlangıç ve bitiş pozisyonlarını da vermektedir. Stanford NER (Manning et al., 2014), İngilizce dilindeki en yaygın kullanılan Varlık Tanımlama uygulamalarından biridir.

Varlık Bağlama (VB), Tanımlı Varlık anmalarını Bilgi Tabanı'ndaki ilgili varlıklarla bağlayarak bu hedefe katkı vermektedir (Han et al., 2011). Şekil 2.2 de görülen örnekte yapısal olmayan metin içinde kısmi olarak bulunan "Mustafa Kemal" ve "Fransa" atıflar referans Bilgi Tabanı'ndaki *siege:Mustafa_Kemal_Ataturk* ve *siege:Fransa* varlıklarıyla bağlanmıştır. Kısmi olarak bulunma sebebi varlığın tam adıyla değil de bir kelimesiyle metinde yer almasından dolayıdır.

[7] http://rtw.ml.cmu.edu/



Şekil 2.1: Varlık Bağlama örneği

Verilen metnin kapsamına göre “Thomas” ünlü bir futbolcu olan “Thomas Müller” ya da “Thomas Edison” varlıklarını gösterebilirken, “oyun” kelimesi futbol oyunu ya da tiyatro oyunu anlamlarına gelebilmektedir. Varlık Bağlama yöntemlerinde metin içindeki atıf bilgi tabanındaki varlıkla söz dizimsel olarak tam bir eşleşmeye sahip olmayabilirken kelimenin anlam ile tam olarak eşleşmesi gerekmektedir (Moro et al., 2014a).

Tanımlı Varlık Çözümleme metin üzerindeki atıfın Wikipedia kaynağından elde edilen sözlük ile çözümlemesi amaçlarken Vikifikasyon metindeki kelime parçalarının kapsama göre en uygun Wikipedia sayfasıyla bağlanmasını amaçlamaktadır. Bu yaklaşım seçilen Wikipedia sayfalarının kendi aralarındaki global benzerliği yok sayarak sadece yerel özelliğe dayanmaktadır. Yerel tutarlılığa dayanan yaklaşımlar her bir atıfı verilen cümlede ayrı ayrı çözümlerken global tutarlılığa dayanan çalışmalar verilen cümledeki kapsam ve alan bütünlüğü içinde bütün atıfların birbirleriyle uyumlu çözümlenmesini amaçlar. Örneğin, verilen cümlede “Michael Jordan” atıfı bilimadamı olan yerine basketbolcuyu etiketliyorsa aynı cümledeki “Bulls” atıfı da Wikipedia kaynağındaki sayfalar arasındaki bağ yapısından benzerlik hesaplanarak “Chicago Bulls” basket takımıyla bağlanır (Ratinov et al., 2011).

Kelime Anlamı Çözümlemesi (KAÇ), kapsam içinde bir kelimenin doğru anlamını bulmayı amaçlar (Navigli, 2009). Kelime çözümlemesi, Doğal Dil İşleme (NLP) alanlarından Makine Çeviri’nin önemli görevlerinden biri olarak 1940’lı yılların sonlarında çalışılmıştır. Bu dönemdeki araştırmacılar hedef kelimenin metin içindeki anlamını istatiksel bilgi, kapsam yada sözlük gibi bilgi kaynaklarından faydalanmanın gerekliliğini önceden farketmişlerdir. Daha sonra 60’lı yıllarda zor bir alan olduğu anlaşılmış 70’li yıllarda ise Yapay Zeka alanındaki yaklaşımların kullanılması kelimenin anlamını çıkarıldığı dili anlamak hedeflenmiştir. Geniş

ölçekli bilgi kaynaklarının olmamasından dolayı Yapay Zeka yaklaşımlarından daha genel sonuçların alınması zor olduğu görülmüştür. 80'li ve 90'lı yıllardan günümüze kadar geniş kapsamlı ve makine okunur dil bilimsel bilgi kaynakları ortaya çıkması (Miller, 1995) ve kitlesel olasılık yöntemlerinin kullanılması Kelime Anlamı Çözümleme için yeni fırsatlar doğurmuştur.

Varlık Bağlama'nın amacı yapısal olmayan metin içindeki atıfın referans alınan Bilgi Tabanı'ndaki en uygun varlıkla bağlanarak keşfedilmesidir. Burada atıf Kelime Anlamı Çözümleme'den farklı olarak anlam karmaşasına yol açacak şekilde kısmi olarak görülebilir. Bu şekilde atıf metnin kapsamına göre farklı anlamlara gelebilir. Kelime Anlamı Çözümleme de ise atıf-metnin içindeki tek kelime yada kelime grubu-referans aldığı sözlükteki anlam ile bire bir eşleşmelidir (Moro et al., 2014b). KAÇ ve VB kullandıkları bilgi kaynağı (sözlük ya da Bilgi Tabanı) ve atıfın kısmi yada birebir eşleşme durumlarına göre iki temel farklılığa sahiptir.

Babelfy (Moro et al., 2014b) çok dil destekli kelime ve tanımlı varlık çözümleme için BabelNet (Navigli and Ponzetto, 2012a) kaynağını kullanmaktadır. BabelNet büyük ölçekli çok dil desteği veren ansiklopedik sözlük ve anlamsal ağ altyapısı olarak WordNet, Open Multilingual WordNet ve Wiktionary faydalanırken; tanımlı varlık isimlerinin bağlanmasında Vikipedi, Wikidata kaynaklarından yararlanmaktadır. YAGO3 (Mahdisoltani et al., 2013) bilgi tabanı YAGO (Suchanek et al., 2007) bilgi tabanının çok dil destekli hali olarak WordNet ve Vikipedi kaynaklarından faydalanmıştır. BabelNet ile aynı hedefe sahipken bilgi tabanı oluşum sürecinde WordNet ve Vikipedi kullanım şekillerinde farklılık göstermektedirler. UWN (de Melo and Weikum, 2012) ve YAGO2 (Hoffart et al., 2011a) bilgi tabanları YAGO3 gelişim sürecine katkı veren önceki çalışmalarıdır. YAGO3, İngilizce için yaklaşık üç buçuk milyon varlık içerirken BabelNet de beş buçuk milyon varlık bulunmaktadır. İki bilgi tabanı da günümüzdeki en güncel, en geniş kapsamlı, alan bağımsız ve çok dil desteği veren örnekler olarak hem kelime anlamı çözümleme hem de varlık bağlamadaki kullanım durumları incelenmiştir.

2.1 Varlık Bağlama Adımları

Varlık Bağlama yöntemleri ağırlık olarak genel kapsamlı Bilgi Tabanı kaynaklarını kullanmaktadır. Navigli (Navigli and Ponzetto, 2012b) manifestosunda alan bağımlı varlık bağlama yöntemlerinin hem başarı hem de keskinlik açısından ileriki yıllarda ön plana çıkacağından bahsetmiştir. Bu açıdan bakıldığında bu tez kapsamında alan bağımlı bilgi tabanı geliştirme sürecinin alan bağımlı varlık bağlama için önemli bir adım olarak görülmektedir.

Alan bağımlı varlık bağlama yönteminde atıf tespiti yapıldıktan sonra aday varlık oluşturmak için atıf-varlık benzerliği ve aday varlık sıralaması için varlık-varlık ilintililiği hesaplanması olmak üzere iki alt yöntemden oluşmaktadır. Benzerlik ve ilintilik arasındaki farka örnek vermek gerekirse “Facebook” ve “Google” varlıkları ontolojideki şirket sınıfında olduklarından yüksek benzerlik ve ilintililik değerlerine sahiptirler. Ancak “Facebook” ve “Mark Zuckerberg” varlıkları birbirlerine benzerlikleri düşük olsa da ilintililik değerleri yüksektir (Cochez et al., 2017). Bu örneğin de gösterdiği gibi birden fazla cümlede sıklıkla aynı bağlamda geçen varlıkların ilintililiği yüksektir. Aşağıdaki şekilde genel olarak alan bağımlı Varlık Bağlama boru hattının adımları gösterilmiştir.

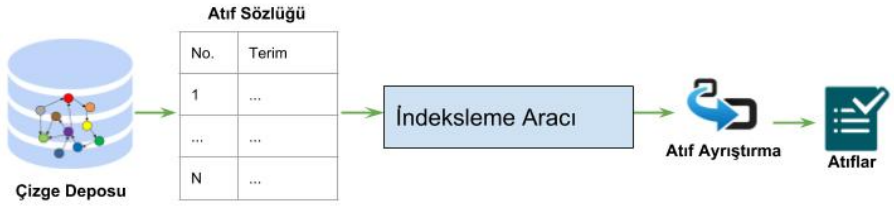
Şekil 2.2: Varlık Bağlama yönteminin temel adımları

Belirli bir alana bağlı metinler ilk önce atıfların tespit edilmesi adımına gelmektedir. Bu adım, sonraki alt bölümde Atıf-Varlık Benzerliği adımı içinde değerlendirilecektir. Bu tez çalışmasının özgün yöntemleri Varlık-Varlık ilintililiği adımında gösterilmiştir. Son adımda ise anlam karmaşıklığı çözümlenen atıfların alan bağımlı bilgi tabanındaki varlıklarla bağlanmış etiketli metin elde edilmektedir.

2.1.1 Atıf tespiti

Atıf tespiti yapısal olmayan metinlerin yapısallaştırılması için öncelikli temel bir adımdır. Bu adımın başarısı diğer adımların başarısını da doğrudan etkilemektedir. Bu sebeple kritik olan atıf tespiti için var olan etiketleme araçlarının kullanılması yada isim sözlüklerinden yararlanılması olmak üzere iki ana yaklaşım mevcuttur.

İngilizce gibi doğal dil işleme alanında yoğunlukla çalışılan diller için Stanford NER (Manning et al., 2014) ve benzeri araçlar kullanılmaktadır. Ancak doğal dil alanında açık kaynak kütüphanesi yada aracı olmayan Türkçe gibi diller için ise Vikipedi yada benzeri kaynaklardan çekilmiş kelime ve kelime gruplarından indeksli sözlükler oluşturulması ile Varlık Tanımla aracına benzer bir sorgulama yöntemi geliştirilmektedir.



Şekil 2.3: Atıf tespitinin genel yapısı.

Şekil 2.3 üzerinde gösterildiği gibi alana bağlı bilgi tabanındaki örnekleri (instance) alfabetik sırada etiketleme sözlüğünde saklanır. Daha sonra bu sözlükten hızlı arama yapılabilmesi için Apache Solr[8] aracı yada Elastic Search aracı[9] ile indekslenir. Verilen metin kelimelere ayrılarak en fazla altılı kelime gruplarında (TagMe çalışması örnek alınarak) n-gram listesi oluşturulur. Son adımda oluşturulan n-gram listesi indekslenmiş atıflar sözlüğünde sorgulanarak atıfların etiketlenmesi sağlanır.

Atıfları etiketlemek için TaGMe çalışmasındaki yöntem kullanılan önemli yöntemlerden biridir. Bu yöntem girdi metnini kelimelere ayırarak en fazla altı ardışık kelimenin yan yana dizilişi ile kelime kombinasyonları elde edilmektedir. Daha sonra her bir kelime yada kelime grubu ontoloji örneklerinden elde edilen sözlükte sorgulanmıştır. Ontolojideki örneklerin daha hızlı sorgulanabilmesi için bütün varlık sözlüğü Elasticsearch[10] aracı ile indekslenmiştir.

Herhangi bir alt dizi probleminde yada bir atıfın sözlükteki başka bir atıfta bulunması durumunda yine TaGMe çalışmasında gösterilen dizilerin sınırlarının belirlenmesi yöntemi kullanılmıştır. Bu yöntemle $m_1 vem_2$ olmak üzere iki atıfı ele alarak m_1 atıfının m_2 atıfındaki bir alt dizi olduğunu varsayalım. Eğer m_2 atıfına ait bağ-olasılığı (lp) $lp(m_1)$ bağ olasılığından büyükse m_1 atıfı mevcut atıflar sözlüğünden silinir. Bu durumda $lp(m_i)$ değeri m_1 atıfına ait bağların öğrenme metinlerinde geçen bütün m_1 sıklıklarına bölünmesi ile bulunmaktadır.

[8] http://lucene.apache.org/solr/
[9] https://www.elastic.co/
[10] https://www.elastic.co/

İsim yada atıf sözlüğü anahtar-değer veri modelini kullanmaktadır. Anahtar olarak atıfın metinde görünür halini tutmaktadır. Değer verileri için Vikipedi bilgi kaynağı kullanılırsa buradaki anlam çözümleme sayfaları ve yönlendirme sayfalarındaki atıfa ait farklı görünme halleri tutulmaktadır. Bu yöntem geleneksel Vikifikasyon yöntemlerinin temelini oluşturmaktadır (Bunescu and Paşca, 2006).

2.1.2 Atıf-varlık benzerliği

Aday varlıkların oluşturulması için önemli bir adım olan atıf-varlık benzerliği için literatürde Varlık Tanımlama araçları ve indekslenmiş sözlüklerin sorgulanması olmak üzere iki alt adım izlenmektedir. Bu tez çalışmasında da bu yönteme dayalı atıf-varlık benzerliği hesaplaması yapılmaktadır. Kelime grupları eşleşirken en uzun dizi eşleşmesi örnek alınmaktadır. Örneğin, "Wicker", "Park" ve "Wicker Park" atıfları içinden karakter boyu en uzun olan ve diğer iki kelimeyi de içeren "Wicker Park" atıfı seçilmektedir. Girdi metninde etiketlenmiş her bir atıf için aday varlıklar çıkarılarak bu varlıklardan atıfa bağlam açısından en benzeri bulunmaktadır.

Varlık popülerliği aday varlıkların oluşturulması aşamasında çok bilinen bir yöntemdir. Bazı varlıklara metinlerde çok fazla karşılaşılmaktadır. Bu durumda bu metinlerde doğal olarak bu varlıklara ait atıflara da sıklıkla karşılaşılmaktadır. Böylece popüler varlık ve atıflar metinlerde tekrar görülme olasılığı yüksek olmaktadır.

$$p(e_i, m_k) = \frac{count_{m_k}(e_i)}{\sum_{e_j \in E} count_{m_k}(e_j)} \tag{2.1}$$

Yukarıdaki formülde e_i aday varlığı için m_k atıfı üzerinden popülerlik özelliği çıkarılmaktadır. Burada $count_{m_k}(e_i)$ fonksiyonu m_k atıfı için hangi sıklıkla e_j varlığı ilişkilendirildiğini saymaktadır.

En benzer atıf-varlık çiftini bulmak için varlıklara ait Wikipedia sayfalarındaki içeriklerden oluşturulan kelime sıklıkları ve bu kelimelerin sıklık sayıları dikkate alınarak Ağırlıklı Jaccard ölçümü kullanılmaktadır. Aşağıdaki formülde atıfın bulunduğu girdi metninden oluşturulan kelime sıklık listesi L_m ile aday varlığa ait Wikipedia sayfasından elde dilen sıklık listesi L_e için Ağırlıklı Jaccard (Weighted Jaccard) hesaplanmaktadır. Sırasıyla atıf ve varlık listelerinin ağırlıkları w_m, w_e her bir listedeki kelime sıklıklarının toplam sıklığa bölünmesi ile elde edilmektedir. Bu değerlerin kesişimlerinin birleşime bölünmesi ile 0-1

aralığında bir değer her bir atıf-varlık listesi için hesaplanmaktır.

$$WJ(L_m, L_e) = \frac{w_m L_m \cap w_e L_e}{w_m L_m \cup w_e L_e} \tag{2.2}$$

Bundan sonraki aşamada, 1 değerine en yakın atıf-varlık çifti belirlenmektedir. Belirlenen girdi metnindeki her bir atıf-varlık çiftine ait varlıklar arasındaki global ilintililik hesapları bir sonraki alt bölümde detaylandırılmaktadır.

2.1.3 Varlık-varlık ilintililiği

Yerel tutarlılığı kullanan Cucerzan, Wikipedia sayfalarındaki kelime kümesinden oluşan kapsam bilgisini de çözümleme sürecine dahil etmiştir (Cucerzan, 2007b). Milne ve Witten (Milne and Witten, 2008) daha sonra bağlanacak varlıklara ait Wikipedia sayfalarının birbiriyle olan bağlantı yapısından hesaplanan anlamsal benzerliği ön plana çıkararak aday kelime anlamının çözümlemesini incelemiştir. Bu yöntem önceden çözümlenmiş kelime anlamlarının kapsam içinde bulunabilirliğine ve anlamını bulmaya amaçlanan hedef kelimeye ne kadar çok bağlantı yapıldığına dayandığı için performansı başarılı olmayabilmektedir. Kulkarni vd. (Kulkarni et al., 2009) ise daha sonra sadece yerel ya da sadece global tutarlılığa dayanan yaklaşımlarda karşılaşılan sorunların üstesinden gelmeyi amaçlamıştır. Alan bağımlı popülerlik hesaplaması veya uzun metin için çözülmesi NP-tam zorluğuna gelebilecek formülünden kaynaklanan problemlere sahiptir.

Milne ve Witten 2008 yılındaki çalışmalarında Anlamsal İlintililik yerel benzerliğe dayanarak bir atıf için belirlenen aday varlıkların kendi aralarındaki ilintililiği hesaplamaktadır.

$$ER(A, B) = \frac{log(max(A, B)) - log(A \cap B)}{log(W) - log(min(A, B))} \tag{2.3}$$

Yukarıda gösterilen eşitlikte, a ve b varlıklarına ait Wikipedia sayfalarının anlamsal ilintililiği ölçülmektedir. Bu iki Wikipedia sayfasına gelen bütün bağlantılar sırasıyla A ve B listeleri olarak verilmektedir. Wikipedia kaynağında yer alan bütün sayfalar da W listesi olarak gösterilmektedir. Anlamsal ilintiliği hesaplamalarında kullanan Han ve arkadaşlarının çalışması (2011) verilen metindeki aday varlıkların kendi aralarındaki tutarlığı ölçmektedir.

Kelime gömme ve anlamsal gömme modellerinin gelişmesi ile birlikte varlıklara ait oluşturulan vektörel temsillerini karşılaştırılarak varlık-varlık ilintililiği hesaplanmaktadır. Varlıklara ait satır satır varlık-ilişki dizileri çıkarıldıktan sonra iki varlığın anlamsal ilintililiği Softmax fonksiyonu kullanılarak bir varlığın diğer varlığın bağlamında olma olasılığı olarak hesaplanmaktadır. Bu şekilde ilintililiği 1'e en yakın çıkan varlık çiftinin birbirine en çok ilintili olduğu anlaşılmaktadır.

$$p(e_0|e_i) = \frac{exp(v_{e0}^{'T} v_{ei})}{\sum_{e=1}^{V} exp(v_e^{'T} v_{ei})} \tag{2.4}$$

Yukarıdaki formül Softmax fonksiyonunun varlık ilintililik hesabı için özelleşmiş hali olarak v_e girdi vektörüne ve $v_e^{'}$ çıktı vektörüne sahiptir. Bu durumda V ise bütün varlıklara ait sözlüğü belirtmektedir.

DoSeR (Zwicklbauer et al., 2016a) çalışmasının RDF2Vec modeline adaptasyonundan önce TaGMe (Ferragina and Scaiella, 2010) çalışmasındaki atıf belirleme yöntemi kullanılarak atıflar işaretlenmiştir. Daha sonra her bir atıf için aday varlıklar elde edilerek bu adaylar için varlık ilintililik hesaplanmıştır. PageRank algoritmasına dayanan yapısında verilen çizgedeki bağların ağırlıklarını varlık geçiş olasılığı (ETP) ile hesaplanmaktadır.

$$ETP(e_u^i, e_v^j) = \frac{cos(vec(e_u^i), vec(e_v^j))}{\sum_{k \in (V \setminus V_i)} cos(vec(e_u^i), vec(k))} \tag{2.5}$$

ETP değeri bir düğümden komşu düğüme ilerleme olasılığını bulmaktadır. Bu olasılıkları bulmak için mevcut durumdaki anlamsal benzerlik hesabı normalize edilmiştir. İki varlık arasındaki anlamsal benzerlik, iki varlığa ait anlamsal izdüşüm vektörleri olan $vec(e_u^i)$ ve $vec(e_v^j)$ üzerinden kosinüs (cos) benzerliği ile hesaplanmaktadır.

2.1.4 NIL varlık problemi

NIL Varlık problemi metinlerde etiketlenen atıflara ait varlık düğümünün verilen bilgi tabanında olmaması durumudur. Bu durum Varlık Bağlama algoritmalarında çok önemli bir görevdir ve bilgi tabanı kaynağının geniş kapsamlı ve tam olup olmaması ile ilgilidir. Bu tez kapsamında olduğu gibi birçok çalışma

doğru hedef varlık olduğunu varsaymaktadır ve verilen bilgi tabanında sürekli olarak bulunduğunu kabul etmektedir.

Literatür taraması yapıldığında bu sorunu çözmek için farklı yaklaşımlar denenmiştir. Aday varlıkların çıkarılmasında akla ilk gelen yöntem atıfa karşılık eğer varlık bilgi tabanında bulunmuyorsa bu düğümü NIL olarak atanmasıdır (Chen et al., 2010). Ancak bu yöntem kesin ve bütün eşleme durumunda geçerli olabilmektedir. Ama parçalı bir isim eşleme durumu gerçekleşiyorsa bu durumda hangi düğümün NIL olacağı net olmayacağı için NIL sayısı artmış olacaktır. Aday varlıkların sıralanması için bazı çalışmalar NIL hedef değeri koyarak bağlanamayacak atıf için bir eşik değeri belirtmektedir (Li et al., 2013). Ayrıca bu eşik değeri için öğreticili öğrenme yöntemlerinden SVM kullanarak atıf-varlık eşleme için kullanan yöntemler de vardır (Ratinov et al., 2011).

NIL probleminin çözümü için bazı yöntemler bu durumu doğrudan anlamsal çözümleme aşaması ile bütünleştirmektedir (Dredze et al., 2010). Başka bir çalışma da belirli bir aday öğeye atıfta bulunan bir yüzey formu için, adayın dil modeli tarafından bu yüzey formunun üretilme olasılığı tarafından üretilen bu yüzey formunun olasılığından önemli ölçüde daha yüksek olmasını varsaymaktadır (Shen et al., 2015). Böylece geliştirilen genel bir dil modeli ile temel olarak, Varlık Bağlama yöntemine altta yatan tarafa bir özellik ekler. Ayrıca bu model bilgi tabanı ve genel diline göre atıfları oluşturduğunu varsaymaktadır. Genel olarak amaç atıfın bağlanıp bağlanamayacağı bir varlık olasılığının hesaplanmasıdır.

NIL probleminin çözümü için farklı yaklaşımlar bu sorunu çözmek için önerilmiştir ancak değerlendirilen veri kümelerine bağlı olarak sonuçlar her zaman ikna edici olmamaktadır. Bu yüzden bu problemi daha fazla iyileştirmek için ek çalışmalar yapılmalıdır.

2.2 Varlık Bağlama Değerlendirme

Bu bölümde şimdiye kadar Varlık Bağlama yöntemi için kritik olan bilgi tabanları ve bunlara dayanan algoritmaların temel adımları detaylandırılmıştır. Gerçekleştirilen bu algoritmaların gerçek dünya uygulamalarında kullanılabilmesi için değerlendirmelerden başarıyla geçmesi gerekmektedir. Bu açıdan bakıldığında bu alt bölümde öncelikle gerçek dünya durumlarına uygun ortamı sağlayan değerlendirme veri kümeleri incelenerek değerlendirme metrikleriyle birlikte varlık etiketleme aracı incelenecektir.

2.2.1 Değerlendime veri kümeleri

Varlık Bağlama sistemlerinin karşılaştırılması için veri kümeleri genel olarak elle yada otomatik olmak üzere iki şekilde elde edilmektedir. Elle elde edilen veri kümeleri elle etiketlenmenin yapılmasından dolayı bu da çok fazla vakit harcanan bir işlem olduğu için genelde küçük boyutlarda metinler içermektedir. Bu tez çalışmasının ana hedefi alan bağımlı Varlık Bağlama sistemlerinin analizi olarak belirlendiği için değerlendirme veri kümelerini alana özgü yada alan bağımsız olarak iki ana kategoride incelenmiştir.

Alan bağımsız veri kümelerinden ilki ACE 2004 (Mitchell et al., 2005) elle oluşturulmuş bir veri kümesi olarak farklı alanlardan 253 atıfın etiketlendiği sadece 57 haber metnini içermektedir. CONLL (Tjong Kim Sang and De Meulder, 2003) yine haber metinlerinden elde edilmiştir. AGDISTIS ve DoSeR çalışmaları bu veri setini deney ortamında kullanmıştır.

Genel bilgi kaynaklarının internet üzerinde gelişmesiyle otomatik olarak veri kümesi üretme işlemine olanak sağlamıştır. Wikilinks (Singh et al., 2012) ve Spitkovsky ve Chang (Spitkovsky and Chang, 2012) çalışmaları Vikipedi üzerinden otomatik veri kümesi oluşturan öncü çalışmalardır. Wikilinks İngilizce dili için otomatik veri kümesi oluşturma metodolojisi gösterirken çok dil desteği sunmamaktadır. Spitkovsky ve Chang (Spitkovsky and Chang, 2012) çalışmasında çok dil destekli bir yaklaşım ile otomatik etiketli veri kaynaklarını üretebilmektedir. Ancak bu iki çalışmada Vikipedi anlam ayrımı sayfalarını kullanarak anlam karmaşıklığının ayarlanması yapmamışlardır.

Li ve arkadaşları (Li et al., 2012) diller arası Varlık Bağlama yöntemleri için çok dil destekli dil kaynağı oluşturan bir yöntem sunmuşlardır. Dil kaynağı oluştururken tanımlı varlıkların etiketlenmesinde kalite standartı olan anlam karmaşıklığı ve çeşitliliği kavramlarını göz önünde bulundurmuşlardır. Anlam karmaşıklığı bir atıfın birden fazla aday varlığa etiketlenmesi durumunda ortaya çıkarken çeşitlilik kavramı aynı atıfa ait metinde geçebilecek birden fazla isim çeşitliliğini göstermektedir. Ayrıca Li ve arkadaşları yeniden yönlendirme ve anlam ayrımı sayfalarını kullanması ve anlam karmaşıklığı seviyesini belirlemişlerdir.

Metin Analizi Konferansları (TAC)[11] Bilgi Tabanı Üretimi (KBP) olmak üzere belirli kategorilerde düzenlenmektedir. KBP kategorisinde yapısal olmayan metinlerden bilgi tabanlarının üretilmesi amaçlanmaktadır. Bu kategoride Varlık

[11] https://tac.nist.gov/2017/

Keşfi ve Bağlanması (EDL) görevi ile tanımlı varlıkların atıflara bağlanması ve bu varlıkların insan, organizasyon yada yer gibi tiplerinin belirlenmesi ile ilgilenmektedir. Bu görevin ana amacı Varlık Bağlama ve etiketleme çalışmarının yüksek anlam karmaşıklığı içeren alan bağımsız veri kümelerinde ve farklı dillerde karşılaştırılmasının yapılmasını ve ekipler arasında fikir alışverişinin sağlanmasını yapmaktır. Bu sebeple, farklı diller ve farklı alanlar için yeterli seviyede anlam karmaşıklığı içeren veri kümelerinin oluşturulması önem kazanmaktadır.

Son olarak GERBIL aracı (Usbeck et al., 2015) çevrimiçi veya çevrimdışı varlık etiketleme çalışmalarının karşılıklı değerlendirmesini yapabilmek için bünyesinde alan bağımsız Varlık Bağlama ve tanımlı varlık çıkarımı yapan araçlar çalışır halde bulunmaktadır. GERBIL aracının ilham aldığı çalışmada (Cornolti et al., 2013) yer alan farklı deney setlerine örnek olarak bilgi tabanına çözümleme (D2KB) ve bilgi tabanına etiketleme (A2KB) gösterilebilinir. Bu tez çalışmasında bu iki seti de hem atıfların etiketlenmesi hemde çözümlenmesi aşamalarında ilgilendirmektedir. Bünyesinde entegre halde bulunan Varlık Bağlama sistemlerine ek olarak GERBIL aracında hali hazırda alan bağımsız ACE 2004 ve CONLL gibi birçok veri kümesi de bütünleştirilmiştir. Aynı zaman kullanıcı tarafından üretilmiş veri kümeleri ve varlık etiketleme sistemleri de mevcut araca entegre edilebilmektedir. Ancak, bu araçta alana özgü ve popüler olmayan bir dile özgü açık veri kümeleri bulunmamaktadır. Bu tez çalışmasının bir diğer amacı da alana ve dile özgü veri kümeleri üreterek alan bağımlı Varlık Bağlama yöntemlerinde kullanılmasını sağlamaktır.

Alan bağımlı veri kümeleri özellikle biyoinformatik alanında çokça rastlanmakla birlikte tamamen varlık etiketlenmesi yerine anlamsal ilintililik yada benzerlik ile ilgili daha özelleştirilmiş durumlar için bulunmaktadır. Pedersen çalışmasında (Pedersen et al., 2007) 29 biyoinformatik kavramının elle yapılmış anlamsal ilintililik için özelleştirilmiş bir veri kümesi bulunmaktadır. Buna ek olarak UMLS (Pakhomov et al., 2010) veri kümesinde 566 tıp terimine ait yine elle hesaplanmış anlamsal benzerlik değerlerini içermektedir.

Örnekleri az olsa da bilgi teknolojileri alanına özgü Bitter Corpus (Arcan et al., 2014) veri kümesi Linux işletim sistemleri için hazrlanmış kullanım kılavuzlarında etiketlenmiş 628 italyanca ve İngilizce terimleri içerirken 637 tane de iki dille ilgili alana özgü kavramları barındırmaktadır. Biyoinformatik çalışmaları dışında diğer alanlara özelleştirilmiş açık kaynak halinde yayınlanan ve doğrudan varlık etiketlenmesi için sunulmuş veri kümeleri bizim bildiğimiz kadarıyla bulunmamaktadır. Bununla birlikte diller arasında paralel (Steinberger et al., 2006) yada iki dil içinde aynı şekilde barındırdığı etiketleme veri kümeleri

(Pamay et al., 2015) ve tamamen doğal dil işleme alanına özel biçimsel yapıları içeren (Sak et al., 2008) veri setleri haricinde literatürde Türkçe diline özgü açık veri kümesi bildiğimiz kadarıyla bulunmamaktadır.

2.2.2 Değerlendirme ölçütleri

Varlık Bağlama ve diğer bir çok metin işlemeye dayanan sınıflandırma yöntemleri için bir karışıklık matrisi (confusion matrix) tanımlanmıştır. Bu matris gerçek değerlerin bilindiği bir test verisi kümesinde bir sınıflandırma modelinin performansını tanımlamak için sıklıkla kullanılan bir tablodur. İkili yada daha fazla sınıflandırıcı içeren örnek bir karışıklık matrisi genel olarak gerçek ve algoritma tarafından tahmin edilmiş doğru ve yanlış sayılarını içermektedir.

Bu matrisin tam sayılar olarak bulunan en temel terimleri:

- **Doğru Pozitifler (TP):** tahmin edilen ile gerçek verinin kesiştiği doğru durumudur.
- **Doğru Negatifler (TN):** tahmin edilen ile gerçek verinin kesiştiği yanlış durumudur. Gerçek pozitiflerle benzerdir ancak burada doğruluk yerine yanlışlık başarılı bir şekilde tespit edilmiştir.
- **Yanlış Pozitifler (FP):** tahmin edilen aslında gerçek veride doğru değildir.
- **Yanlış Negatifler (FN):** Olmadığı tahmin edilen durum aslında gerçekte vardır.

Bu temel terimlerden hareketle doğruluk (accuracy), hassasiyet (precision), anma (recall) oranları bulunmaktadır. Doğruluk sınıflandırıcının tespit ettiği başarılı sonuçların toplam tespit edilenlere bölünmesidir. Varlık Bağlama açısından bakıldığında bu oran yöntemimizin ne kadar başarılı bir şekilde atıf ve varlık çiftlerini eşleştirdiğini göstermektedir.

$$A = \frac{TP + TN}{\sum_e M(m, e)} \tag{2.6}$$

Varlık Bağlama yöntemi için gerçek negatiflerin olmasına gerek yoktur çünkü olmayan bir atıf-varlık eşlemesinin doğrulunu kontrol edilmeyebilir. Bu sebeple

paydada sadece gerçek pozitiflerin toplam eşlemelere bölünmesi ile A doğruluk oranına erişilmektedir.

$$R = \frac{TP}{FN + TP} \tag{2.7}$$

Anma R oranı gerçekte doğru olanın hangi sıklıkla başarılı bir şekilde tahmin edildiğini göstermektedir. Buradaki doğru bulunan eşlemeler anma oranını ve dolayısla yöntemin başarısını doğrudan etkilemektedir. Burada kritik olan yanlış negatiflerin sayıca az olması yöntemin genel başarısı için önemli bir etkendir.

$$P = \frac{TP}{FP + TP} \tag{2.8}$$

Hassasiyet P oranı gerçekte doğru olmayan eşlemenin önem taşıdığı doğru bulunan eşlemelerle oranlanmasındır. Anmadan farklı olarak bu ölçümde yanlış pozitiflerin önemli bir rol oynamaktadır. Burada kritik olan yanlış pozitiflerin sayıca az olması yöntemin genel başarısı için önemli bir etkendir. Son olarak anma ve hassasiyet oranlarının ağırlıklı harmonik ortalaması ile $F1$ ölçümü bulunmaktadır.

$$F1 = \frac{2(PR)}{(P + R)} \tag{2.9}$$

Buradan 0 ve 1 aralığındaki bir değer ile Varlık Bağlama yönteminin genel başarısı izlenebilmektedir. Böylece 1 değerine yakın olan yöntemlerin daha başarılı olduğu gözlenmektedir.

2.3 Bilgi Tabanları

Bilgi tabanı herhangi bir bilgi tabanı kullanan Varlık Bağlama yönteminin yapı taşıdır. Yapı taşı olarak varlık hedef sözlüğünü tanımlar ve farklı türde varlıklarla ilgili bilgileri sağlamaktadır. Genel olarak bir varlık açıklama üzerinden yönelimli olarak tanımlanmış yada örneklerle ve kullanım durumu ile uzamsal olarak belirtilmiştir. Yönelimli tanımlar eşanlamlılar sözlüğü yada bir varlığın mantıksal temsili olabileceği gibi çizge tabanlı bilgi tabanlarında örneğin RDF olarak da sunulmaktadır. Genişlemeli tanımlar ise bir varlığın varlık etiketli metinlerde sunulduğu gibi kullanım bağlamı hakkında bilgi vermektedir.

2.3.1 Genel amaçlı bilgi tabanları

Bilgi tabanı kullanan Varlık Bağlama yaklaşımlarında bilgi kaynaklarının kapsamı ve tutarlılığı genel performansı artırmak ve etkin çözüm üretebilmek adına önem kazanmaktadır. Bu açıdan bakıldığında geleneksel bilgi kaynakları genel olarak WordNet (Miller, 1995) ve Vikipedi kaynaklarından oluşmaktadır:

- **WordNet :** WordNet doğal dil işleme alanında popüler olan İngilizce için sözlüklendirilmiş bir veri kümesidir. Ontoloji bakış açısıyla yaklaşıldığında her bir sınıf yani kavrama ait eş anlam kümesi (synset) bulundurmaktadır. Bu anlam kümeleri aynı anlamı taşıyan kelime listelerini içermektedir. Bu listedeki her kelime isim, zarf yada sıfat gibi tip bilgisi de gösterilmektedir. Ayrıca, eş anlam kümelerindeki terimler çok anlamlı kelimelerden de oluşabilmektedir. WordNet her anlam kümesi için kısa bir tanımlama cümlesi de barındırmaktadır. İlişkiler açısından bakıldığında genelleme (hypernymy) ve özelleştirme (hyponymy) ilişkisi olan taksonominin en temel yapıtaşı Is-A ilişkisi ağırlık kazanmaktadır. Buna benzer 12 farklı ilişki yapısı da barındırmaktadır.

- **Vikipedi :** Vikipedi gönüllü editörler tarafından oluşturulmakta olan çok dil destekli çevrimiçi bir ansiklopedidir. Mevcut olarak Vikipedi sadece İngilizce dilinde 44 milyon makale barındırmaktadır. Her bir Vikipedi makalesi herhangi bir tanımlı varlık yada kavram için detaylı bilgi sunmaktadır. Vikipedi sayfalarında tablolar, bilgi kutuları, geri yönlendirme, anlam ayrımı, kategori yada diller arası olmak üzere çeşitli bağlantılar içeren büyük bir Vikipedi bağ yapısı içermektedir. Geri yönlendirme sayfalarında aynı anlama gelen kavramlar için birden fazla bağlantı içerirken anlam ayrımı sayfalarında bir kavramın birden fazla farklı sayfasına yönlenme imkanı bulunmaktadır. Vikipedi kategori sayfaları da kavram ve varlıklar hakkında ilgili oldukları alan detaylı şekilde verilmektedir.

Yukarıda detaylandırılan bilgi kaynaklarından melezlenerek farklı inşa desenleri ile geliştirilen bilgi kaynaklarından en bilinenleri şu şekilde listelenmiştir:

- **DBpedia (Bizer et al., 2009):** DBpedia çok dil destekli bilgi tabanı olarak yarı yapısal Vikipedi bilgi kaynağından bilgi kutusu, kategori, anlam ayrımı ve geri yönlendirme bağlantılarını kullanarak geliştirilmektedir. Yapısallaştırıldığı için Sparql sorgu dili ile sorgulanabilmektedir ve Veri

Web'i bulutunun çekirdeği konumunda bulunmaktadır. WordNet içindeki taksonomik düzene ek olarak varlıklar hakkında insan, organizasyon ve yer gibi tip ilişkilerini ve ayrıca çok iyi bir şekilde detaylandırılmış sinema filmi yıldızı yada futbolcu gibi sınıflandırma bilgilerini de barındırmaktadır.

- **YAGO (Suchanek et al., 2007):** YAGO çalışmasında WordNet taksonomi bilgisi üst veri modelinin çekirdeğini oluştururken Vikipedi üzerinden çekilen verilerle de gerçeklerin çoğaltılması sağlanmaktadır. YAGO modelinde her şey tanımlı varlık olarak saklanmakta ve insanlar, organizasyonlar, yerler, kitaplar ve şarkılar gibi birçok örneği içermektedir. Kavramlar ve Is-A ilişki hiyerarşisi WordNet üzerinden gelerek Vikipedi kategori sayfalarını kullanan sezgisel yöntemle iki bilgi kaynağını birleştirilmesi sağlanmıştır.

- **BabelNet (Navigli and Ponzetto, 2012a):** BabelNet bilgi tabanı yine WordNet ve Vikipedi kaynaklarının melezlenmesi ile oluşturulmuş bir dilbilgisel ve çok dil destekli yapısal sözlüktür. BabelNet modelindeki melezleme YAGO çalışmasından farklı olarak çizgeye benzemektedir. Varlıklar arasındaki taksonomik yapı belirgin bir şekilde gösterilmemekle birlikte yine Is-A ilişkilerini WordNet üzerindeki eş anlam kümelerine dayanarak almaktadır. Eş anlam kümelerini merkezine alarak çok dil desteği sağlama amacıyla makine çevirim yöntemi sayesinde farklı dillere çevrilmiş kavramlar kelimeler torbası (BoW) ve çizge tabanlı yöntem ile WordNet ve Vikipedi yer alan aynı kavram ve varlıklar eşleştirilmiştir.

- **Knowledge Vault (Dong et al., 2014) :** Bu çalışmada ana prensip çekirdek olarak ele alınan Freebase (Bollacker et al., 2008) bilgi tabanını bütün Webteki sayfalarındaki bilgileri çıkararak çekirdek kaynaktaki kavram ve varlıkları zenginleştirmeyi amaçlamaktadır. Çıkarılmış web dokümanlarını, web tablolarını yada etiketlenmiş metinleri öğreticili yöntemlerle mevcut bilgi tabanına entegre edilmesini amaçlanmaktadır. Mevcut bulunan Freebase kaynağındaki sınıflandırma ve taksonomik altyapıyı kullanarak bilgileri çekilmiş web kaynaklarındaki gerçeklerin doğrulanması yapılmaktadır. Bu doğrulama işlemi için her bir gerçek üçlü için güven değerlerini dikkate almaktadır.

Yukarıdaki listede bilgi tabanı çalışmalarından çıkarılan desenlerden Knowledge Vault modelinde sunulan bilgi tabanının web kaynakları ile zenginleştirildiği desen bu tez çalışmasında örnek alınmıştır. Seçilen alanda var olan Bilgi Tabanı belirlenmiş ve daha sonra Vikipedia ve alana özgü veb kaynakları gibi harici kaynaklarla zenginleştirilmiştir. İlk önce bu desen kapsamında alana bağımlı sinema, müzik, vd. gibi ontolojiler bulunmuştur. Daha sonra mevcut

ontoloji yapısallığı veb sitelerinden daha gelişmiş olan Vikipedia kaynağı ile zenginleştirilmiştir. Wikipedia kaynağında yer alan bilgi kutuları, yönlendirme ve çözümleme bağlantıları ile hem ontolojideki sınıflar geliştirilmiş hem de örnekler çekilmiştir.

2.3.2 Alana özgü bilgi tabanları

Genel amaçlı bilgi tabanlarının aksine bazı veri yoğun görevlerde alana özgü bilgi tabanlarının geliştirilmesinin önü açılmıştır. Genel amaçlı bilgi tabanları geniş kapsamlı olmalarına rağmen özellikle biyoinformatik alanında yoğun bir şekilde bulunan detaylı ve karmaşık bilgi yoğunluğu için yeterli gelememektedir. Bu yüzden biyoinformatik ve son zamanlarda bilimsel çalışmalar alanında da ağırlıklı önem kazanmış bilgi tabanlarına örnekleri aşağıda sıralanmıştır:

- **DBLP (Ley, 2002)** Dijital Kaynakça ve Kütüphane Projesi çevrimiçi bir sistem olarak bilgisayar bilimi üzerine dergi makaleleri, konferans çalışmaları ve diğer yayınları içermektedir. Genel olarak bu bilgi tabanı makaleler, yazarlar, konferansların düzenlendiği yerleri, anahtar kelimeleri ve yayınların yıllarını saklamaktadır. Güncel haliyle DBLP bilgi tabanı bir milyonun üzerinde yazar, yaklaşık 2.6 milyon yayın ve 7000 adet konferans yer bilgisini barındırmaktadır. Güncel olarak yayınların Bibtex bilgileri sayesinde Google Scholar [12] ile de bağlantı sağlanmaktadır.

- **KnowLife (Ernst et al., 2015)** Biyoinformatik alanında bir çok bilgi tabanı mevcuttur. Bunlardan en bilinenlerinden biri olarak KnowLife biyomedikal bilgi tabanı olan UMLS (Bodenreider, 2004) (Unified Medical Language System) genişletilerek kurulmuştur. Sadece sağlık alanındaki bilimsel yayınlar için değil, ansiklopedik sağlık portalları ve çevrimiçi topluluklar için de kapsamlı deneyler yürütülerek farklı konfigürasyonlara dayanan bir bilgi tabanı oluşturulmuştur. En iyi yapılandırılmış haliyle KnowLife sağlık alanında genleri, semptomları, tedavileri ve çevresel ve yaşam tarzı risk faktörlerini kapsayan 13 ilişki için $\%93$'lük bir doğrulukla 500.000'den fazla gerçek içerir. KnowLife, farklı internet kaynaklarından otomatik olarak oluşturulan sağlık ve yaşam bilimleri için geniş bir bilgi tabanı olarak yerini almaktadir.

- **LinkedMDB (Hassanzadeh and Consens, 2009)** Sinema alanına özgü bir bilgi tabanı olarak IMDb[13] sitesinden çözümlenerek yapısallaştırılmıştır.

[12] https://scholar.google.com.tr/
[13] https://www.imdb.com/

Örnek zengini bir bilgi tabanı olarak daha hafif bir ontoloji yapısı vardır. Temel olarak tip bilgisini tutmaktadır. LinkedMDB projesi, birkaç önemli film web kaynağını birleştiren ilk açık bağlantılı veri kümesinin bir gösterimini sağlar. LinkedMDB'nin maruz kaldığı veritabanı, büyüyen Linking Açık Veri bulutunun bir parçası olan mevcut web veri kaynaklarına ve IMDb gibi popüler hareketli web sayfalarına yüzlerce bin RDF bağlantısı içeren milyonlarca RDF üçlüsünü içeriyor. LinkedMDB, bağlantıları bulmak için en son birleştirme teknikleri kullanarak ve bağlantıların kalitesi ve bunları türetmek için kullanılan teknikler hakkında ek RDF üst veri modellerini sağlayarak çok miktarda ve kaliteli bağlantılar oluşturmanın ve korumanın bir yolunu sunmaktadır.

Bu bilgi tabanlarının her birini kendi alanları için örnek olarak gösterilmiştir. Bu alanlarda daha çok çalışma mevcuttur ancak tez kapsamında genel bir izlenim sağlamak açısından bu üç bilgi tabanı örneği seçilmiştir. Daha az varlık veya gerçek sağlayan diğer birkaç alana özgü bilgi tabanları sınırlı sayıda varlık barındırdığı ve zorunluluk nedeniyle Varlık Bağlama çalışmalarında pek kullanılmamaktadır. Bilgi tabanı olarak yeterli olabilmeleri için temel varlık bilgisini tam olarak kullanılmak üzere Varlık Bağlama algoritmasında güçlü bir şekilde uymaası gerekmektedir. Farklı bir bilgi türünden farklı bir şekilde nasıl yararlanılacağı açık bir sorundur ve kapsam bilgisinin Varlık Bağlama algoritmasında daha verimli kullanılması için genişletilmesi önemli bir araştırma alanıdır. Sonraki alt bölümde bilgi tabanı kullanan Varlık Bağlama yönteminin temelinde yatan adımları bu bakış açısından incelenecektir.

2.4 Sonuç ve Değerlendirme

Bu bölümde, önceki bölümdeki Varlık Bağlama probleminin tanımlanması ve tez içeriğinin genel yapısının verilmesinden sonra yöntemin temel yapısını ve değerlendirme sürecinin detayı verilmiştir.

Varlık Bağlama yöntemlerinin bilgi tabanı kullanan algoritmalarında genel amaçlı yada alana özgü bilgi tabanlarının tanıtılması ve kullanım durumları açıklanmıştır. Daha sonra temel adımları metinde geçen atıfların belirlenmesi ve bu atıfların verilen bilgi tabanındaki referans varlık ile eşlenmesi sırasında geçen çözümleme aşaması incelenmiştir. Bu aşamada atıf ile varlık arasındaki benzerlik ile birlikte varlık ile diğer aday varlıkların belirli özellikleri doğrultusunda sıralanması adımları detaylandırılmıştır.

3 LİTERATÜR ÇALIŞMALARI

Varlık Bağlama son on yılda kapsamlı bir şekilde incelenmiştir. Bu süre içinde Farklı görev tanımları gelişmiş ve çeşitli zorluklar ortaya çıkmıştır. Bu bölümde, Varlık bağlama ve onun temelleri daha spesifik olarak, farklı problemleri içeren Varlık Bağlama alanına genel bir bakış açısı ile ele alınmıştır.

Varlık Bağlama ile ilgili temel kavramlar ve yöntemler ele alınacaktır. Sonrasında belirtilen yaklaşımlar ve yöntemlere göre geliştirilen sistemler incelenerek literatürdeki durum karşılaştırmalı olarak ortaya konulacaktır.

Sonraki alt bölümlerde sırasıyla ortaya çıkan formülasyonlar, zorluklar ve ilgili görevler, Varlık veri tabanları olarak kullanılan tipik bilgi tabanları ve algoritmaların nasıl değerlendirildiğini gösterilmiştir.

3.1 Varlık Bağlama Yöntemleri

Bilgi Çıkarımı, Bilgi Alma, Makine Çevirisi gibi önemli araştırma alanları kelimelerdeki belirsizlikleri çözmekten büyük yarar sağlar. Dahası, Doğal Dil İşleme gibi farklı araştırma toplulukları, Semantik Web ve Veri Madenciliği topluluğu, kelime belirsizliği sorununu ele almıştır. Farklı şekillerde çerçevelenmiş olarak Varlık Bağlama yerine Kelime Anlamı Çözümleme, Çapraz Belge Eş-Referans Çözünürlük ve Kayıt Bağlantı isimleri ile de anılmaktadır. Araştırma toplulukları çok sayıda algoritma ve yaklaşım geliştirdiği ancak şu ana kadar dört görevi ayrı ayrı ele almıştır.

Varlık Bağlama sistemleri öğreticili, öğreticisiz ve Bilgi Tabanı yöntemleri olmak üzere üç ana yaklaşım ile tanımlı varlıkların çözümlenmesini amaçlamaktadır (Shen et al., 2015). Bu yöntemlerinin artı ve eksileri özetlenerek Bilgi Tabanı seçiminin sebepleri ve diğer yöntemlerin Bilgi Tabanı'na yapabileceği katkılar incelenmiştir.

- **Öğreticili Öğrenme Yöntemleri:** Elle etiketlenmiş öğretici veri kümesinden Makine Öğrenme tekniklerini kullanarak sınıflandırıcı öğrenmeyi amaçlar. Sınıflandırmayı kendisine ait cümle içindeki kelimenin uygun anlamı gibi pozitif ve negatif örnekler etiketleyerek öğrenmektedir (Alpaydin, 2014). El ile etiketlenme gereksinimi yüzünden Bilgi Edinimi Darboğazı (Gale et al., 1992) problemiyle karşı karşıya gelinmektedir. Çok büyük miktarda

etiketlenmiş veriye ihtiyaç duymasından dolayı yeni dil ve alanlar eklendikçe bu veri etiketleme sürecinin tekrarlanması zorlaşmaktadır (Zhong and Ng, 2010).

- **Öğreticisiz Öğrenme Yöntemleri:** Öğreticili Öğrenme'ye göre performansı daha düşük olsa da el ile etiketlenme ihtiyacı duymayan, Korpus üzerinden kapsama göre yada kelimelerin vektörel komşuluklarındaki benzerliğe bakarak kümelenmesidir. Derlem işlenmemiş ham halde bulunabilmektedir. Ham Korpus örneği olarak bir milyon kelimeden oluşan metin veri kümesi olan Brown Corpus (Francis and Kucera, 1982) ve yaklaşık 30 milyon kelime içeren metinlere sahip Wall Street Journal Corpus (Charniak et al., 2000) verilebilir. Korpus ham olabileceği gibi kelime anlamı etiketlenmiş şekilde de bulunabilir. SemCor (Miller et al., 1993) en çok kullanılan Anlam-Etiketli Korpus olarak 352 metin üzerinde 234 bin anlam etiketi barındırmaktadır. Kapsam ve kelime olmak üzere iki yolla kümelenme çalışılmıştır. Büyük ölçekli Korpus ihtiyacı ve bu ihtiyaç karşılansa bile sözlük gibi evrensel kabul gören kelime anlam bilgisine dayanmaması gibi dezavantajları bulunmamaktadır. Veri dağınıklığı probleminin yaşanması ve kıyaslamanın öznel olabilmesi gibi sorunlarla karşılaşılmaktadır (Agirre and Soroa, 2009).
- **Bilgi Tabanı Yöntemleri:** Bilgi zenginliği olan (Bilgi Tabanı, sözlük) yada Bilgi zenginliği olmayan (Korpus) olmak üzere iki farklı yaklaşıma ayrılır. BabelNet, YAGO ve UWN gibi bütünleştirilerek zenginleştirilen Bilgi Tabanı ve sözlük kaynaklarının varlığı ve bu kaynaklarda varlıkların birbirleriyle yapısal bir ilişki halinde bulunması kelime ve varlık çözümleme için kolaylık sağlamaktadır. Bu tez çalışmasında Bilgi Tabanı kaynakları Kelime Anlamı Çözümleme (KAÇ) ile bütünleştirilmiş Bilgi Tabanı üzerinde çalışılacağından KAÇ destekli yada desteksiz Varlık Bağlama yaklaşımları detaylandırılmıştır.

J. Han ve arkadaşlarının çalışmasına göre tipik bir Varlık Bağlama sistemi aşağıdaki aşamalardan oluşmaktadır (Shen et al., 2015):

- **Aday Varlıkların Oluşturulma Aşaması (AVO):** Her bir varlık atıfı için Bilgi Tabanı'ndaki olası varlıkların listelenmesidir. Bu aşamada ilgili olmayan varlıkların elenip gerekli olanların seçilmesi önem kazanmaktadır. Uygulanılan teknikler; sözlük tabanlı, girilen metin üzerinden yüzeysel ismin (surface form) çıkarılması, arama motorları üzerinde uygulanan teknikler mevcuttur.

- **Aday Varlıkların Sıralanma Aşaması (AVS):**Mevcut bağlam içinde seçilmiş aday varlıklardan hangisinin verilen kasıta en uygun olduğunun belirlenmesidir. Öğreticiyle ve öğreticisiz öğrenme teknikleri kullanılarak aday varlıklar sıralanır. Daha sonra en yüksek değeri alan aday varlık mevcut anma için en uygun varlık olarak seçilir. Kelime veya Tanımlı Varlık çözümleme süreci bu aşamada gerçekleştirilir.

İngilizce dışındaki yaygın kullanımı olmayan dillerde hali hazırda tanımlı varlık çıkarması yapabilecek araçlar[14] olmadığı için ilk önce Tablo 3.1 de görüldüğü gibi anmaların belirleme aşamasının sütun olarak eklenmesi gerekmektedir.

Atıf Tespiti aşaması için genel olarak Stanford NER aracı tercih edilmekle birlikte Vikipedi kaynaklı İsim Sözlüğü'nden yararlanan araçlar da olmuştur. Veb sorgu logları, indeksleme, tek yada çok kelime gruplarından oluşan metinsel parçacıklar, N-gram analizi ve K-Shingle gibi kelimelerin sonraki yada önceki birlikte bulunma durumları incelenerek de metin üzerindeki kasıt (anma) belirlenmiştir. Aday Varlıkların Oluşturulma aşamasında İsim Sözlüğü, popülerlik ve kapsam benzerliği, yerel veya global özellikler gibi Vikipedi kaynaklı yaklaşımlara ek olarak Çizge tabanlı anlamsal benzerlik bulan yaklaşımlar da bulunmaktadır. Aday Varlıkların Sıralanması aşamasında ise ilk iki aşamada sunulan aday varlıkların metindeki kapsamına göre Öğreticili, Öğreticisiz veya olasılık yöntemleri kullanılarak sıralanması çalışılmıştır. Çizge Tabanlı yaklaşımlarda Kasıt-Varlık çizgesi oluşturup adayların sıralanması ön plana çıkan çalışmalarda görülmüştür. Bu kategorilendirmeye ek olarak bilgi tabanı bağımsız ve bağımlı yöntemler olarak iki ayrı tablo halinde Varlık Bağlama sistemleri verilmiştir.

3.1.1 Bilgi tabanı bağımlı sistemler

Bilgi Tabanı'na bağımlı çalışmalar sadece seçilen bilgi tabanı üzerinde çalışabilirken bağımsız çalışmalar ise herhangi bir bilgi tabanı üzerinde de başarılı sonuçlar verebilmektedir.

Geniş ölçekli Bilgi Tabanı kaynaklarının ortaya çıkmasıyla Varlık Bağlama fikri doğmuş ve Tanımlı Varlık Çözümleme yada Vikifikasyon gibi görevlere yeni fırsatların çıkmasına olanak sağlamıştır. Varlık Bağlama, Tanımlı Varlık Çözümleme ve Vikifikasyon olmak üzere iki alt görevden oluşur. Tanımlı Varlık Çözümleme (Cucerzan, 2007b) metin üzerindeki kasıtın bilgi tabanı ile

[14]https://github.com/AKSW/FOX

Çizelge 3.1: Bilgi Tabanı bağımlı Varlık Bağlama sistemleri.

Sistem	**Atıf Tespiti**	**AVO**	**AVS**
Cucerzan (2007)	Veb Sorgu Logları	Global özellikler	Öğreticisiz Öğrenme
Wikify (2007)	İsim Sözlüğü	Olasılık Yöntemleri	Öğreticisiz Öğrenme
Wikipedia Miner (2008)	Atıf Belirleme	Anlamsal İlgi	Öğreticili Öğrenme
TagMe (2010)	Apache Lucene	İsim Sözlüğü	Öğreticili Öğrenme
AIDA (2011)	Stanford NER	Kapsam Benzerliği	Çizge Tabanlı
Illinois Wikifier (2011)	İsim Sözlüğü	Yerel ve Global Özellikler	Öğreticili Öğrenme
Linden (2012)	İsim Sözlüğü	İsim Sözlüğü	Sıralamalı Öğrenme
TagMe 2 (2012)	İsim Sözlüğü	Ortak kararlı global özellikler	Çizge Tabanlı
Dexter (2013)	İsim Sözlüğü	İsim Sözlüğü	Çizge Tabanlı
Tulip (2014)	SolrTextTagger	Vikipedi Konu Kategorileri	Konu Benzerliği
Dexter 2 (2014)	Shingle Çıkarsama	İsim Sözlüğü	Çizge Tabanlı
Babelfy (2014)	Metinsel Parçalar	Çizge Tabanlı	Çizge Tabanlı
AIDA-light (2014)	İsim Sözlüğü	Kolay Varlık Bulma	Çizge Tabanlı
WAT (2014)	Open NLP	Olasılık Yöntemleri	Atıf-Varlık Çizgesi
Kea (2016)	N-gram Analizi	İsim Sözlüğü	Önceden kapsam belirleme

bağlanmasını amaçlarken Vikifikasyon metindeki kelime parçalarının kapsama göre en uygun Vikipedi sayfasıyla bağlanmasını amaçlamaktadır. Bu yaklaşım seçilen Vikipedi sayfalarının kendi aralarındaki global tutarlılığı yok sayarak sadece yerel özelliğe dayanmaktadır.

Bu problemin üstesinden gelebilmek için Cucerzan kelimelere ait dilbilgisel kapsam bilgisini de çözümleme sürecine dahil ederek global bir yaklaşım önermiştir (Cucerzan, 2007b). Milne ve Witten daha sonra anlamsal benzerliği ön plana çıkararak anlam karmaşıklığı olmayan tek anlamlı kelimeler kapsamı içinde aday kelime anlamının çözümlemesini incelemiştir (Milne and Witten, 2008). Bu yöntem önceden çözümlenmiş kelime anlamlarının kapsam içinde bulunabilirliğine ve anlamını bulmaya amaçlanan hedef kelimeye ne kadar çok bağlantı yağıldığına dayandığı için performansı başarılı olmayabilmektedir. Kulkarni et. al. ise

daha sonra sadece yerel yada sadece global özelliklere dayanan yaklaşımlarda karşılaşılan sorunların üstesinden gelmeyi amaçlamıştır (Kulkarni et al., 2009). Alan bağımlı popülerlik hesaplaması veya uzun metin için çözülmesi NP-tam zorluğuna gelebilecek formülünden kaynaklanan problemlere sahiptir. Illinois Wikifier (Ratinov et al., 2011) çıkarılan atıflar için giriş metnini Wikipedia çapalarını ve başlıklarını kullanarak kendi geliştirdiği Varlık Tanımlama sistemi çıkarmaktadır. Toplu olmayı amaçlayan bir optimizasyon problemi olarak tüm ifadeler arasında tutarlılığa bakılarak varlık çözümleme yapılır.

Atıf tespiti için isim sözlüğü kullanan çalışmalardan LINDEN (Shen et al., 2012) Vikipedi'de gömülü zengin anlambilimsel bilgiden ve bilgi tabanının sınıflandırmadan yararlanarak, Vikipedi ve WordNet bilgi kaynaklarını birleştiren bir bilgi tabanıyla metin içinde adlandırılmış varlıkları bağlamıştır. Bu zaman aralığında TagMe çalışmasının geliştirilmiş versiyonu olarak sunulan TagMe 2 (Ferragina and Scaiella, 2012) çalışmasında önceki sürümdeki varlık çözümleme aşaması oylama şeması ile zenginleştilerek en yüksek oyu alan yöntem seçilmektedir. Vikipedi bağ yapısını kullanarak varlık ilintililiği yöntemine (Milne and Witten, 2008) dayanmaktadır.

Çizge tabanlı varlık çözümlemeye dayanan Dexter (Ceccarelli et al., 2013) bir çerçeve olarak popüler algoritmaları içeren bir araçtır. Daha sonra indeksleme aracını kullanarak atıfları tespit eden Tulip (Lipczak et al., 2014) sisteminin amacı bir belgede varlıkları tespit etmek ve söz konusu bahsi geçen Freebase ile bağlamaktır. Bunu başarmak için belgelerin içeriğini yakalayan ilgili varlıkların çekirdek alt kümesine odaklanan varlık adayları kümesini hazırlamaktadır. İlişki kuvveti, varlık özelliklerinden elde edilen bir benzerlik olarak ölçülür. Her varlık, Wikipedia'nın 120 dil sürümünden alınan bilgilere dayanarak oluşturulmuş bir kategori grafiğinden çıkarılan doğru ve kompakt özellikli bir vektörle temsil edilmektedir.

Öğreticili Öğrenme yöntemlerinde karşılaşılan Bilgi Çıkarımı probleminin üstesinden gelebilmek için geniş ölçekli ansiklopedik ve dil bilimsel bilgi kaynaklarının yapısal bilgi kaynaklarına bütünleştirmesi amaçlanmıştır. Babelfy ve AIDA-light (Nguyen et al., 2014b) uygulamalarında görüldüğü gibi mevcut durumda çalışmalar önce geniş kapsamlı büyük bir Bilgi Tabanı elde ettikten sonra verilen girdi metnine göre bu Bilgi Tabanı'nı daraltarak en yoğun alt çizgeden çözümlemeyi yapmayı amaçlamaktadırlar. WAT (Piccinno and Ferragina, 2014) TagMe sisteminin yeni bir sürümü olarak toplu varlık çözümlemesi için çizge tabanlı yaklaşımlar kullanmaktadır. Üstelik, bu yaklaşımlar, söz konusu metindeki tüm söz varlığı çiftlerinin tutarlılığını vurgulayan küresel tutarlılık

yaklaşımlarına odaklanmaktadır. Son olarak Kea (Waitelonis and Sack, 2016) aday varlıkları çıkarmak için metnin n-gramlarının DBpedia ile eşleştirmektedir. Ön işlem aşamasında ilk önce atılan tweet metinlerini temizler ve normalleştirmektedir. Daha sonra ağırlık grafik ölçümleri, bağlı bileşen analizi, varlıkların merkezi ve yoğunluk gözlemleri varlıkları çözümleme aşamasında kullanılmaktadır.

3.1.2 Bilgi tabanı bağımsız sistemler

Alan bağımsız Varlık Bağlama sistemlerinden DBpedia Spotlight (Mendes et al., 2011) ve Babelfy (Moro et al., 2014a) global tutarlılık yaklaşımlarına dayanan ve önemli başarım değerleri alan çalışmalardır. Ancak Vikipedi bağ yapısından bağımsız sistemlerin gereksinimi son yıllarda ortaya çıkmıştır. Bilgi Tabanı bağımsız yaklaşımlar da Çizelge 3.3 listelenmiştir.

Çizelge 3.2: Bilgi Tabanı bağımsız Varlık Bağlama sistemleri

Sistem	**Yıl**	**Atıf Tespiti**	**AVO**	**AVS**
AGDISTIS	2014	FOX	Anlamsal Benzerlik	HITS
DoSeR	2016	Stanford NER	Anlamsal İlintililik	PageRank

AGDISTIS (Usbeck et al., 2014) Vikipedi bağ yapısından bağımsız olan bir sistem olarak Stanford NER aracnı kullanarak atıfları belirlemektedir. Belirlenen atıflar için aday varlıkların çıkarılması ile verilen bilgi tabanındaki çözümleme alt çizgesi edilmektedir. Bu alt çizgeye aday varlıklar arasındaki benzerliğe ve HITS algoritmasına dayanan global tutarlılık hesabı ile en uygun atıf-varlık çiftini eşleştirmektedir. Global tutarlılığa dayanan çalışmalar son yıllarda anlamsal izdüşüme dayanan yöntemleri benimsemeye başlamıştır. Anlamsal izdüşüm, kavramların çok boyutlu vektörleri olarak kosinüs benzerliği ile hesaplanan bir benzerliği tanımlamak için kullanılmaktadır. Bu yöntem sadece kelime içeren metinlerde Word2Vec modeli olarak çalışılmıştır (Mikolov et al., 2013a). DoSeR çalışmasında ise kelimeler yerine bilgi tabanlarındaki sınıf, ilişki ve örneklerin sorgulanması ile elde edilen dizileri kullanan anlamsal izdüşüm modeli oluşturulmaktadır. Daha sonra bu model PageRank algoritmasındaki bağların ağırlıklarını hesaplamak için varlıklar arasındaki anlamsal benzerliği bulmak için kullanılmıştır. DoSeR çalışması mevcut çalışmaların başarımını yüzde 10 civarında geliştirdiği için bu tez çalışmasındaki hedef yaklaşım olarak ele alınmıştır. Bu açıdan bakıldığında anlamsal izdüşümlerin başka bir yöntemi RDF2Vec (Ristoski and Paulheim, 2016) kullanılarak farklı ontoloji yapılarından elde edilen RDF izdüşümlerinin seçilen DoSeR ve AGDISTIS çalışmalarındaki etkisi gözlenmiştir.

Varlık Bağlama sistemlerindeki temel yapı olan bilgi tabanlarını anlamsal

gömme için elverişli şekilde ontoloji yapılarını seçmek önem kazanmıştır. Di Noia ve arkadaşları (Noia et al., 2016) bilişim teknolojileri alanında anlamsal ilintililik çizgesi oluşturmuştur. Bu çalışmada bilişim teknolojilerinde seçilen kök listesi yazılım araçları ve donanımsal parçalardan oluşan varlıklardır. Daha sonra DBpedia ile zenginleştirilerek anlamsal arama ile anahtar kelime ile aramadan daha etkin bir çözüm sunmayı hedeflemişlerdir. Etkin çözümü de varlıklar arasındaki anlamsal ilintililik hesabı sayesinde başarmışlardır.

Çizge çekirdekleri (de Vries and de Rooij, 2015) de bilgi tabanından verimli varlık-ilişki dizileri çıkararak anlamsal izdüşüm hesaplamasına katkı vermektedir. Bu tür çalışmalar mevcut ontolojiyi çift yönlü çizgelere ayırmaktadırlar. Daha sonra bu çizge içinde alt ağaçlar oluşturmaktadırlar. Bu alt ağaçlar içinde birbirine en yakın olanları varlık-ilişki dizilerine dönüştürmektedirler. Böylece hesaplama süresini en aza indirerek anlamsal izdüşüm başarımı yükseltmeyi hedeflemektedirler.

Varlık Bağlama çalışmaları genellikle varlıkları açıklığa kavuşturmak için Wikipedia bağlantı yapısına odaklanır. (Cucerzan, 2007b). Dahası, bu sistemlerden bazıları (Mendes et al., 2011) büyük ölçüde Vikipedi kaynağından türetilen bilgiye dayanmaktadır. DBpedia (Mendes et al., 2011) ve YAGO2 (Hoffart et al., 2011a) gibi üsler. DBpedia Spotlight (Mendes et al., 2011), DBpedia'yı temel bilgi tabanı olarak kullanır ve vektör uzay modeline ve kosinüs benzerliğine bağlıdır. AIDA (Hoffart et al., 2011b) YAGO2'ye dayanır ve bir varlığın önceki olasılığını birleştirir, İşletmenin adayları arasındaki söz ve tutarlılık arasındaki benzerlik, toplu bir varlık sapmasına neden olur. TagMe (Ferragina and Scaiella, 2010), Wikipedia bağlantı yapısına dayalı anlaşma yaklaşımını kullanarak kısa metinlerin açıklamalı annotasyonunu hedefleyen ve Wikipedia algılama bağlantılarını kullanır. WAT (Piccinno and Ferragina, 2014), TagMe'nin geliştirilmiş bir sürümüdür ve toplu varlık anlamsızlığı için çizge tabanlı algoritmalar kullanır. Üstelik, bu yaklaşımlar, söz konusu metindeki tüm söz varlığı çiftlerinin tutarlılığını vurgulayan küresel tutarlılık yaklaşımlarına odaklanmaktadır.

Geleneksel çalışmalar ağırlıklı olarak, Wikipedia ve diğer varlıkları temsil etmek için Vikipedi ile ilişkili el yapımı özellikleri kullanır. Bu çalışmalar, bu yüksek bağımlılık nedeniyle altta yatan bilgi tabanını değiştirdikten sonra daha kötü performans gösterebilir. Bu bağımlılığı azaltmak için, AGDISTIS (Usbeck et al., 2014), tespit edilen ifadeler için aday varlıkları seçer ve bu adaylar için bir belirsizlik grafiği oluşturur. Ardından, çizge tabanlı HITS algoritması, en iyi söz varlığı çiftlerini eşleştirmek için belirsizlik çizgesine uygulanır. Dahası, bilgi gömülmeleri, geniş yapılandırılmamış metinlerden sürekli sözcük vektörleri

temsillerini içeren kelime toplama performansının teşvik edilmesinden sonra da popüler hale gelmiştir (Mikolov et al., 2013a). Wang ve diğ. Ayrıca, Wikipedia bağlantıları ve varlıkları kullanarak kelime ve varlık yerleşimlerini birbirine bağlamak için bilgi yerleşimlerini tanıtın. Doser (Zwicklbauer et al., 2016a), aday varlıkları ayırmak için Kişiselleştirilmiş-PageRank algoritmasının girdisi olarak kelime yerleşimlerini kullanır. Bu çalışmalara benzer şekilde, nöral modellerin bir girdisi olarak RDF embriyonlarından (Ristoski and Paulheim, 2016) faydalandık.

3.1.3 Dizi öğrenmeye dayanan sistemler

Son yıllarda, yapay sinir ağı modelleri el yapımı özellikler olmaksızın daha iyi genellemeyi teşvik etmenin bir yolu olarak sunulmuştur. He ve diğ. (He et al., 2013), Torba-Kelimeler girdisine dayalı bağlam temsili için bir ileriye dönük ağdan yararlanan ön yöntemlerden biridir. Sun ve diğ. cite sun2015modeling, birleşik bir şekilde söz, varlık ve bağlamdaki yerleşimleri kullanarak bir sinir ağı yaklaşımı sunar. Bağlam temsili için bir Konvolüsyel Sinir Ağı modelini kullanırlar ve bağlamsal kelimelerin sözlerin etrafındaki konumlarını düşünürler. Varlık bağlamındaki girdiler ve aday varlıklar arasındaki benzerliği hesaplayan bir sıralama görevi olarak, kurumun belirsizliğini tanımlarlar. Fang ve diğ. (Fang et al., 2016) tren kelimesi ve aynı vektör temsili içindeki varlık yerleşimleri. Bu toplantılara dayanarak, bağlantı eş oluşumları, önceliğe değinme ve varlıkların bağlamsal ilişkileri gibi özellikleri içeren etkili bir anlam çıkarma modeli önermektedirler. Yamada ve diğ. (Yamada et al., 2016) aynı zamanda, sözcükleri ve varlıktaki yerleşimleri, nesneleri birbirinden ayırmak için aynı sürekli vektör uzayına birleştiren bir ortak öğrenme yöntemi sunar. Bu iki çalışmaya benzer olarak Gupta ve ark. Ayrıca, Gupta et al. (Gupta et al., 2017) bağlam, söz ve varlıkların ortak kodlamasını, aday varlıklar için tanımlanmış ince taneli türde bilgilerle genişletir. Benzer şekilde, bu varlık türü bilgisini aday varlıkların filtrelenmesi için bir alan göstergesi olarak kullanırız.

NeuPL (Phan et al., 2017), varlıkları ayırmak için UKHA ve dikkat mekanizması kullanır. Ayrıca, en kolay çiftinden başlayarak sözde-varlık çiftleri eşleşen hızlı bir Çift-Bağlama algoritması sağlar. NeuPL, konum bilgisi ve kelime sıralamasını dikkate alır. Bu nedenle, her bir sözün sol ve sağ taraflarının bağlamını modellemek için iki UKHA ağı kullanılır. Bizim çalışmamız, tüm çiftlerden ziyade daha yakın sayma-varlık çiftlerini çözme açısından NeuPL'ye benzer. Daha yakın sayma-varlıklı çiftlerin ayırma yöntemimiz, NeuPL çift bağlanma yönteminden farklıdır. Bizim yöntemimiz, belirli bir alan için adlandırılmış bir varlık tanıma görevi olarak daha yakın komşuları ayırmak için CRF'leri kullanır.

Çizelge 3.3: Dizi Öğrenmeye dayanan Varlık Bağlama sistemleri

Sistem	**Atıf Tespiti**	**AVO**	**AVS**
(He et al., 2013)	-	Bağlama dayalı dizi modeli	Softmax
(Sun et al., 2015)	Stanford NER	Atıf ve Varlık için ortak dizi modeli	Softmax
(Yamada et al., 2016)	-	kelime ve varlık dizi modeli	Softmax
(Gupta et al., 2017)	Stanford NER	bağlam, kelime, varlık tipi içeren dizi modeli	Softmax
NeuPL (Phan et al., 2017)	-	UKHA	Aday çifti benzerliği

Etki alanına özgü Varlık Bağlaması, bilgi yoğun etki alanlarının performanslarını iyileştirmek için de önemlidir (Navigli, 2013). AIDA-light (Nguyen et al., 2014a), küresel tutarlılığı varlıkları ayırmaya ve YAGO2 ve Vikipedi etki alanı hiyerarşisini sömürmeye karar verir. Pantel ve Fuxman (Pantel and Fuxman, 2011), belirli bir sorgu ile kurumun alaka düzeyini tahmin eden bir ilişkilendirme modeli önermektedir. Bu model, büyük bir ürün kataloğu için sorgu varlık tıklama grafiği ve sorgu tıklama günlüklerini kullanır. Dalvi ve diğ. tweet'lerin ve restoranların haritasını çıkarmak için tweetlerin coğrafi bilgilerini kullanır (Dalvi et al., 2012). D'ouza ve Ng (D'Souza and Ng, 2015) harita hastalığı, biyomedikalde ilgili varlıklar ile klinik raporlarda bahseder. ontoloji. Son zamanlarda, Shen ve ark. (Shen et al., 2018), DBLP [15] ve IMDb IMDb[16]'nin HIN ağındaki yeniden varlık varlıkları içeren bir varlık anlaşılmaması görevi olarak çok-tipi ve birbirine bağlı nesneler de dahil olmak üzere heterojen bilgi ağları (HIN) kullanır. Buna ek olarak, Deola, DBLP'ye (Liu et al., 2016) 'da yazarlar için çevrimiçi bir varlık bağlama sistemi sağlar. Deola, belirsiz aday varlıkları çözmek için PageRank algoritmasını ve yazar, yıl ve yayın başlıklarını içeren meta-yolları kullanır. Zhang ve diğ. (Zhang et al., 2016), adlandırılmış varlıklar olarak tanınmaması gereken ortak ifadeleri ifade eden sahte adlandırılmış varlıklar aracılığıyla alana özel bir varlık bağlama yöntemi önerir. GnoSSEA, yerel bir bilgi tabanı oluşturmaktan ziyade belirli alanların DBPedia'dan çıkarılması nedeniyle bu çalışmadan farklıdır. Ayrıca, GnoSSEA el yapımı özellikleri kullanmak yerine dizi-sıralı modeller uygular.

[15] http://www.dblp.org/
[16] http://www.imdb.com/

3.2 Değerlendime Veri Kümeleri

Varlık Bağlama sistemlerinin karşılaştırılması için veri kümeleri genel olarak elle yada otomatik olmak üzere iki şekilde elde edilmektedir. Elle elde edilen veri kümeleri elle etiketlenmenin yapılmasından dolayı bu da çok fazla vakit harcanan bir işlem olduğu için genelde küçük boyutlarda metinler içermektedir. Bu tez çalışmasının ana hedefi alan bağımlı Varlık Bağlama sistemlerinin analizi olarak belirlendiği için değerlendirme veri kümelerini alana özgü yada alan bağımsız olarak iki ana kategoride incelenmiştir.

3.2.1 Alan bağımsız değerlendirme veri kümeleri

Alan bağımsız veri kümelerinden ilki ACE 2004 (Mitchell et al., 2005) elle oluşturulmuş bir veri kümesi olarak farklı alanlardan 253 atıfın etiketlendiği sadece 57 haber metnini içermektedir. CONLL (Tjong Kim Sang and De Meulder, 2003) yine haber metinlerinden elde edilmiştir. AGDISTIS ve DoSeR çalışmaları bu veri setini deney ortamında kullanmıştır.

Genel bilgi kaynaklarının internet üzerinde gelişmesiyle otomatik olarak veri kümesi üretme işlemine olanak sağlamıştır. Wikilinks (Singh et al., 2012) ve Spitkovsky ve Chang (Spitkovsky and Chang, 2012) çalışmaları Vikipedi üzerinden otomatik veri kümesi oluşturan öncü çalışmalardır. Wikilinks İngilizce dili için otomatik veri kümesi oluşturma metodolojisi gösterirken çok dil desteği sunmamaktadır. Spitkovsky ve Chang (Spitkovsky and Chang, 2012) çalışmasında çok dil destekli bir yaklaşım ile otomatik etiketli veri kaynaklarını üretebilmektedir. Ancak bu iki çalışmada Vikipedi anlam ayrımı sayfalarını kullanarak anlam karmaşıklığının ayarlanması yapmamışlardır.

Li ve arkadaşları (Strassel et al., 2008) diller arası Varlık Bağlama yöntemleri için çok dil destekli dil kaynağı oluşturan bir yöntem sunmuşlardır. Dil kaynağı oluşturutken tanımlı varlıkların etikenlenmesinde kalite standartı olan anlam karmaşıklığı ve çeşitliliği kavramlarını göz önünde bulundurmuşlardır. Anlam karmaşıklığı bir atıfın birden fazla aday varlığa etiketlenmesi durumunda ortaya çıkarken çeşitlilik kavramı aynı atıfa ait metinde geçebilecek birden fazla isim çeşitliliğini göstermektedir. Ayrıca Li ve arkadaşları yeniden yönlendirme ve anlam ayrımı sayfalarını kullanması ve anlam karmaşıklığı seviyesini belirlemişlerdir.

Metin Analizi Konferansları (TAC)[17] Bilgi Tabanı Üretimi (KBP) olmak üzere belirli kategorilerde düzenlenmektedir. KBP kategorisinde yapısal olmayan metinlerden bilgi tabanlarının üretilmesi amaçlanmaktadır. Bu kategoride Varlık Keşfi ve Bağlanması (EDL) görevi ile tanımlı varlıkların atıflara bağlanması ve bu varlıkların insan, organizasyon yada yer gibi tiplerinin belirlenmesi ile ilgilenmektedir. Bu görevin ana amacı Varlık Bağlama ve etiketleme çalışmarının yüksek anlam karmaşıklığı içeren alan bağımsız veri kümelerinde ve farklı dillerde karşılaştırılmasının yapılmasını ve ekipler arasında fikir alışverişinin sağlanmasını yapmaktır. Bu sebeple, farklı diller ve farklı alanlar için yeterli seviyede anlam karmaşıklığı içeren veri kümelerinin oluşturulması önem kazanmaktadır.

Son olarak GERBIL aracı (Usbeck et al., 2015) çevrimiçi veya çevrimdışı varlık etiketleme çalışmalarının karşılıklı değerlendirmesini yapabilmek için bünyesinde alan bağımsız Varlık Bağlama ve tanımlı varlık çıkarımı yapan araçlar çalışır halde bulunmaktadır. GERBIL aracının ilham aldığı çalışmada (Cornolti et al., 2013) yer alan farklı deney setlerine örnek olarak bilgi tabanına çözümleme (D2KB) ve bilgi tabanına etiketleme (A2KB) gösterilebilinir. Bu tez çalışmasında bu iki seti de hem atıfların etiketlenmesi hemde çözümlenmesi aşamalarında ilgilendirmektedir. Bünyesinde entegre halde bulunan Varlık Bağlama sistemlerine ek olarak GERBIL aracında hali hazırda alan bağımsız ACE 2004 ve CONLL gibi birçok veri kümesi de bütünleştirilmiştir. Aynı zaman kullanıcı tarafından üretilmiş veri kümeleri ve varlık etiketleme sistemleri de mevcut araca entegre edilebilmektedir. Ancak, bu araçta alana özgü ve popüler olmayan bir dile özgü açık veri kümeleri bulunmamaktadır. Bu tez çalışmasının bir diğer amacı da alana ve dile özgü veri kümeleri üreterek alan bağımlı Varlık Bağlama yöntemlerinde kullanılmasını sağlamaktır.

3.2.2 Alan bağımlı değerlendirme veri kümeleri

Alan bağımlı veri kümeleri özellikle biyoinfomatik alanında çokça rastlanmakla birlikte tamamen varlık etiketlenmesi yerine anlamsal ilintililik yada benzerlik ile ilgili daha özelleştirilmiş durumlar için bulunmaktadır. Pedersen çalışmasında (Pedersen et al., 2007) 29 biyoinformatik kavramının elle yapılmış anlamsal ilintililik için özelleştirilmiş bir veri kümesi bulunmaktadır. Buna ek olarak UMLS (Pakhomov et al., 2010) veri kümesinde 566 tıp terimine ait yine elle hesaplanmış anlamsal benzerlik değerlerini içermektedir.

Örnekleri az olsa da bilgi teknolojileri alanına özgü Bitter Corpus (Arcan

[17] https://tac.nist.gov/2017/

et al., 2014) veri kümesi Linux işletim sistemleri için hazrlanmış kullanım kılavuzlarında etiketlenmiş 628 italyanca ve İngilizce terimleri içerirken 637 tane de iki dille ilgili alana özgü kavramları barındırmaktadır. Biyoinformatik çalışmaları dışında diğer alanlara özelleştirilmiş açık kaynak halinde yayınlanan ve doğrudan varlık etiketlenmesi için sunulmuş veri kümeleri bizim bildiğimiz kadarıyla bulunmamaktadır. Bununla birlikte diller arasında paralel (Steinberger et al., 2006) yada iki dil içinde aynı şekilde barındırdığı etiketleme veri kümeleri (Pamay et al., 2015) ve tamamen doğal dil işleme alanına özel biçimsel yapıları içeren (Sak et al., 2008) veri setleri haricinde literatürde Türkçe diline özgü açık veri kümesi bildiğimiz kadarıyla bulunmamaktadır.

Elle etiketlenen metinler önyargılı olma eğilimindedir çünkü insanlar genellikle varlık ek açıklamaları için bilinen terimleri seçer. Ayrıca, bu açıklama süreci, popüler olmayan terimler için bazen gürültülüdür. Bu nedenle, Wikipedia, birçok geliştiricinin küratörlüğünü yaptığı ve yapılandırılmış bir ek açıklama sürecini içerdiği için seçilmelidir. MSNBC (Cucerzan, 2007b), IITB (Kulkarni et al., 2009) ve Wikilinks (Singh et al., 2012) genel varlık açıklama görevler için deneysel veri setlerini önermektedir (Singh et al., 2012). Wikilinks, Vikipedi'ye bağlantılar yoluyla otomatik olarak oluşturulmuş büyük ölçekli bir etiketli corpus sağlar. Wikilinks, bir masif koleksiyonunu tanımlamak için otomatik bir yöntem sunuyor varlık miktarları ve emeklemeye dayanır. Vikipedi sayfalarında bağlantıyı ve bağlantı metni anımsatma olarak kullanma. Bununla birlikte, Wikipedia anlam karmaşıklığı sayfalarını kullanmak için belirsizlik düzeylerinin seviyesi için de kullanılabilir ve bu doğrudan Wikilinks'te gösterilmemiştir.

Belirsizlik, belirsiz ve benzersiz varlıklar arasındaki orandır ve varlık annoratörleri için daha gerçekçi bir ortam sağlar (Li et al., 2012). belirli alanlar için açıklamalı metinleri belirsizliği ayarlamak ve oluşturmak için, biz yeni bir çalışmada kullanmak alıntı son Vikipedi, belirli alanlar için İngilizce [18] içinde dökümü ayıklar (Inan and Dikenelli, 2017) . Bunu yapmak için Wikipedia kategorisi sayfalarını ve DBpedia "dct: subject" [19] özelliğini kullanırlar. Ayrıca, seçilen alan adlarında Vikipedi ayırma sayfalarının kullanıldığı belirsiz bir ortam sağlarlar. Bir örnek olarak, söz *Wicker Park* bir Vikipedi anlam ayrımı sayfasıdır [20] vardır ve filmde belirsizliği arttırmak için kullanılabilir domain.

Film değerlendirme veri seti İngilizce olarak 123 açıklamalı metin içerir. Her metin için ortalama varlık sayısı 4.99'dur ve toplamda 614 varlık vardır. Filmler,

[18]https://dumps.wikimedia.org/enwiki/20170420/
[19]http://purl.org/dc/terms/subject
[20]https://en.wikipedia.org/wiki/Wicker_Park

yönetmenler ve başrol oyuncuları, Wikipedia makalelerinin bilgi kutularından çıkarılır ve DBpedia tarafından referans varlıklar ile eşleştirilir. Bu varlıkların netleştirme sayfaları, film alanı için değerlendirme veri kümesindeki belirsizlik oranını artırmak için müzik ve konum gibi diğer etki alanlarında ayıklanır. Değerlendirme veri setinin belirsizlik oranı, tüm belirsiz varlıkların film alanı için çıkarılmış toplam tekil varlık sayısına bölünmesiyle hesaplanan 48,79 % 'dır. Bu nedenle, Varlık Bağlama sistemlerini değerlendirmek için daha gerçekçi belirsiz bir veri kümesi oluşturulabilir.

3.3 Varlık Bağlama Değerlendirme Ölçütleri

Varlık Bağlama'nın değerlendirilmesi için en son yaklaşımları ile GERBIL kıyaslama çerçevesinde gerçekleştirilmiştir. Bilgi tabanındaki ilgili varlıklar için tespit edilen sözlerin anlaşılmaması üzerine odaklanan Bilgi Tabanına (D2KB) göre etiketleme seçilmiştir. Bu görevde, ilgili varlığa eşleştirmek için belirli bir atıf garanti edilir.

Varlık Bağlama görevinde Cornolti ve ark. (Cornolti et al., 2013), standart değerlendirme önlemlerini, söz konusu varlık çiftleri arasındaki doğru eşleşmeleri tanımlamak için daha uygun bir formata genişletir. Örneğin, doğru (TP) ve yanlış pozitif (FP), doğru (TN) ve yanlış negatifler (FN) gibi karışıklık matrisinin elemanları Doğru etiketlemeyi tanımlayan bir ikili ilişki T olarak tanımlanabilir. D için bir giriş belgesi için d için belirlenen sette belirtilen doğru varlıklardır. Test edilen varlık annotator tarafından bulunan gerçek sonuç a olsun. Sonra aşağıdaki tanımları elde edilir:

$$\begin{aligned} TP(a,e,T) &= \{x \in a | \exists y \in e : T(y,x)\} \\ FP(a,e,T) &= \{x \in a |\ \nexists y \in e : T(y,x)\} \\ TN(a,e,T) &= \{x \notin a |\ \nexists y \in e : T(y,x)\} \\ FN(a,e,T) &= \{x \in e |\ \nexists y \in a : T(y,x)\} \end{aligned} \tag{3.1}$$

Denklem 3.1 'ye bağlı olarak, F1 ölçüsü Makro ve Mikro ölçütlere genelleştirilebilir. Makro ölçütler, tüm açıklamalı dokümanlardaki her bir doküman üzerinde karşılık gelen önlemin ortalaması iken, Mikro ölçüler tüm etiketleri birlikte ele alır ve böylece daha fazla etikete sahip belgelere daha fazla önem verir.

$$
\begin{aligned}
Micro_P(A,E,T) &= \frac{\sum_{d\in D}|TP(a_d,e_d,T)|}{\sum_{d\in D}(|TP(a_d,e_d,T)|+|FP(a_d,e_d,T)|)} \\
Micro_R(A,E,T) &= \frac{\sum_{d\in D}|TP(a_d,e_d,T)|}{\sum_{d\in D}(|TP(a_d,e_d,T)|+|FN(a_d,e_d,T)|)} \\
Micro_{F1}(A,E,T) &= \frac{2.Micro_P(A,E,T).Micro_R(A,E,T)}{Micro_P(A,E,T)+Micro_R(A,E,T)} \\
Macro_P(A,E,T) &= \frac{\sum_{d\in D}P(a_d,e_d,T)}{|D|} \\
Macro_R(A,E,T) &= \frac{\sum_{d\in D}R(a_d,e_d,T)}{|D|} \\
Macro_{F1}(A,E,T) &= \frac{2.Macro_P(A,E,T).Macro_R(A,E,T)}{Macro_P(A,E,T)+Macro_R(A,E,T)}
\end{aligned}
\tag{3.2}
$$

3.4 Sonuç ve Değerlendirme

Bu bölümde literatürdeki çalışmalar kronolojik olarak değerlendirilmiş ve son olarak Varlık Bağlama sistemlerinin farklı örnekleri karşılaştırılmıştır. Varlık Bağlama sistemlerinin ilk başlarda Vikipedi bağ yapısına dayalı olarak varlıkları çözümlediği görülmüştür (Milne and Witten, 2008; Cucerzan, 2007b). Bununla birlikte Vikipedi kaynağından türemiş bilgi tabanlarına bağımlı çalışmalar da gözlenmiştir (Mendes et al., 2011; Hoffart et al., 2011b). DBpedia Spotlight (Mendes et al., 2011) sistemi DBpedia bilgi tabanını kullanarak vektörel uzay modeli ve kosinüs benzerlik değerlerine dayanır. AIDA (Hoffart et al., 2011b) sistemi varlıkların popülerliği, atıfların bağlamındaki benzerlik ve aday varlıklar arasındaki tutarlılık değerlerinin toplu varlık çözümleme aşamasına sokmaktadır. TagMe (Ferragina and Scaiella, 2010) Vikipedi çıpa bağ metinlerini atıf tespiti aşamasında kullanırken Vikipedi bağ yapısına dayanan bir anlaşma yöntemi kullanmaktadır. WAT (Piccinno and Ferragina, 2014) sistemi TagMe çalışmasının gelişmiş bir versiyonu olarak çizge tabanlı yöntemleri en yüksek değeri veren sistemi seçen bir oy verme yapısı kullanmaktadır.

Son yıllarda yapay sinir ağı modellerinin gelişmesi ve kelime gömme (Mikolov et al., 2013a) yönteminin ortaya çıkması ile Varlık Bağlama için daha genel ve elle özelliklerin çıkarılmasına gerek duymayan yöntemler sunulmuştur. Bunlardan öncül çalışma olan (He et al., 2013) ileri beslemeli bir yapay sinir ağı modelini varlıkların bağlamında kullanarak çözümleme yapmıştır. Daha sonra (Yamada et al., 2016) kelime ve varlık gömme bilgisini aynı vektörel uzayda barındırarak çözümlemede kullanmıştır. Buna benzer olarak atıf bağlamı,

varlıkların alan bilgisi ve varlıkların bağlamı ile birlikte aynı vektör uzayında harmanlayarak aday varlıkların sıralanması için kullanmıştır (Gupta et al., 2017). Ancak bu yöntemler yatay olarak gittikçe büyüyen çok büyük vektör uzayına sahip oldukları için çözümleme aşaması için büyük bir hesaplama maliyeti doğmaktadır.

Diziden diziye öğrenme modelleri (Sutskever et al., 2014) yatay ve büyük vektör uzayı yerine dikeyde katmanlar arası çözümleme olanağı tanımaktadır. Yapay Sinir Ağı Makine Çeviri yaklaşımında orjinal olarak herhangi bir doğal dilde verilen bir kaynak cümleden şartlı olasılığa dayanan kıyaslamalı dil modeli sayesinde başka bir dildeki hedef cümle tahminlenmektedir. Böylece, bu tez kapsamında kaynak cümlelere atıf dizisi olarak davranarak hedef varlık dizilerinin global olarak çözümlenmesi sağlanmaktadır. Sonuç olarak yatay vektörel uzaya dayanan yöntemlerdeki gibi ekstra bir global çözümleme yöntemine gerek kalmadan diziden diziye öğrenme modeli ile çözüm sunulmaktadır.

Alan yönlendirmesi girdi veri boyutunu düşürmesi sayesinde sinir ağı modellerine dayanan yöntemler için daha verimli bir çözüm sağlamaktadır. D'souza ve Ng (D'Souza and Ng, 2015) tanı atıflarını biyomedikal ontolojideki varlıklarla eşleştirmektedir. Son zamanlarda (Shen et al., 2018) Heterojen Bilgi Ağlarını (HBA) heterogeneous information networks (HIN) çok tipli ve birbiri ile bağlı nesneleri varlık çözümleme aşamasında kullanmıştır. HBA ağını DBLP[21] ve IMDb[22] alan bağımlı bilgi tabanları ile kullanmıştır. Bu çalışmalardan farklı olarak bu tez kapsamında sunulan metod sıfırdan bilgi tabanı kurmak yada belirli alan bağımlı bilgi tabanlarını kullanmak yerine Bağlı Veri bulutunun çekirdeği olan DBpedia bilgi tabanını alanlara ayırararak anlamsal gömme modellerini çıkarmıştır.

[21] http://www.dblp.org/
[22] http://www.imdb.com/

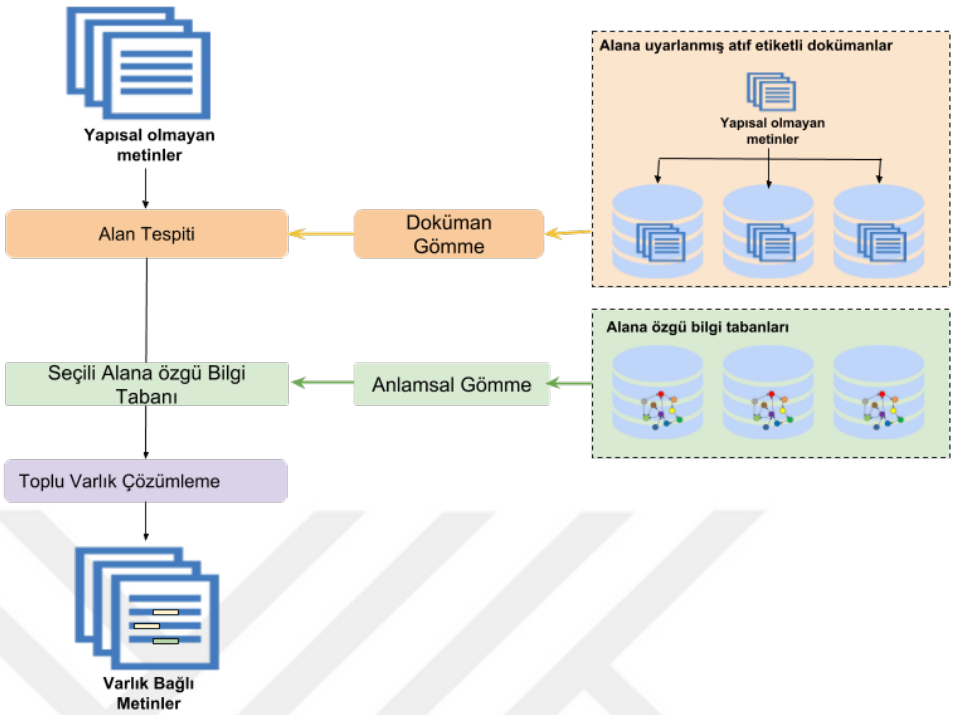
4 YÖNTEM

Daha önceki bölümlerde Varlık Bağlama probleminin tanımı, problemin çözümü için gereken adımlara genel bakış ve literatürdeki çalışmalar incelenmiştir. Bu tez çalışması kapsamında, alana özgü bilgi tabanı kullanarak Varlık Bağlama için bir dizi öğrenme yöntemi önerilecektir.

Varlık Bağlama probleminin temel amacı atıfları etiketli metinler için her bir atıfa ait aday varlıklardan doğru atıf-varlık eşleşmesini yapabilmektir. Bu açıdan bakıldığında varlık çözümleme aşamasında aday varlıkların kendi aralarındaki ilintililik değerlerine göre sıralanması temel adım olarak görülmektedir. Son zamanlardaki Varlık Bağlama çalışmaları global tutarlılık kavramını dikkate almaktadır. Bu kavramın temel hedefi aynı metindeki bütün atıflara ait aday varlıkların metnin bağlamına uygun şekilde birbirleriyle tutarlılıklarını dikkate almaktadır (Milne and Witten, 2008; Kulkarni et al., 2009; Ratinov et al., 2011). Son zamanlardaki çalışmalara benzer olarak (Zwicklbauer et al., 2016a; Usbeck et al., 2014) sunulan Varlık Bağlama yöntemi girdi metinlerinin baştan tespit edilmiş atıflara sahip oldukları varsayılmaktadır. Aynı zamanda bir diğer varsayım ise her bir atıfın bir yada daha fazla aday varlık ile eşleşebilmesidir. Böylece GnoSSEA bilgi tabanı ve doküman kataloğu bağımsız bir alana uyarlanmış yöntem sunmaktadır.

Şekil 4.1 sunulan yöntemin ana mimarisini göstermektedir. İlk adımda yapısal olmayan girdi metinlerinin alan tespiti yapılmaktadır. Bunu yapmak için genel amaçlı herhangi bir doküman kaynağından belli bir alana ait dokümanlar çekilmiştir. Daha sonra bu dokümanlardan her alana ilişkin doküman vektörleri elde edilmektedir. Alanı tespit eldilmek istenen yapısal olmayan metinin vektörü ile alana ilişkin doküman vektörleri karşılaştırılmaktadır. En yüksek benzerliğe bu iki vektöre bakılarak girilen metnin alanı tespit edilmektedir. Bu tez kapsamında doküman gömme modeli olarak Doc2Vec (Le and Mikolov, 2014) modeli seçilmiştir. Verilen girdi metninin alanı tespit edildikten sonra hesaplama maliyetini düşüren verimli bir yöntem sunulmaktadır.

Tezde sunulan yöntemin ikinci aşaması, alanı tespit edilen girdi metinlerinin alanı ile bağlantılı olan alana özgü bilgi tabanının seçimini içermektedir. Tezin kapsamında önkoşul olarak alana uyarlanmış atıf etiketli metinlerdeki atıfların karşılığı olan varlıklara sahip bilgi tabanlarının kullanılması beklenmektedir. Ancak doküman yığını ile bilgi tabanı arasındaki bu atıf ve karşılığı olan varlıkların eşleştiren herhangi bilgi kaynakları bu tezde önerilen yöntemde kullanılabilinir.



Şekil 4.1: Sunulan yöntemin ana mimarisi.

Alana özgü bilgi tabanı seçildikten sonra bu bilgi tabanları için anlamsal gömme modelleri üretilmektedir. Tez kapsamında RDF2Vec (Ristoski and Paulheim, 2016) modeli anlamsal gömme modeli olarak kullanılmaktadır. Kullanılan bilgi tabanı ve doküman yığınlarında el yapımı özelliklerin çıkarılmasına gerek olmamaktadır. Alan tespiti ve bilgi tabanı seçimi adımları doküman ve anlamsal gömme modellerine dayandığından dolayı herhangi bir bilgi kaynağında çalışabilmektedir.

Önerilen çalışmanın son adımı, toplu varlık çözümleme aşamasının alana özgü bir şekilde ele alan bir diziden diziye öğrenme modeline dayanmaktadır (Sutskever et al., 2014). Bu öğrenme modelinin girdisi olarak atıf ve varlıklara ait anlamsal gömme modelleri ve alana özgü doküman modelleri bu tez kapsamında kullanmaktadır. Böylece, varlıkların etkin bir şekilde çözümlenmesi için hesaplama zamanı ve gereksiz verilerin azaltılması sağlanmaktadır. Ayrıca, kaynak atıf dizisinden hedefteki aday varlık dizisinin tahminlenmesi özgün bir toplu varlık çözümleme yöntemi olarak bu tezde sunulmaktadır. Özet olarak bu tezin önerdiği yöntem yeni bir alandan gerekli bilgileri ekleyerek genişleyebilen ölçeklenebilir bir mimari sunmaktadır.

Bu bölümün alt başlıklarında sırasıyla girdi metinlerinin alan tespiti için doküman gömme modülü, alana özgü bilgi tabanlarından elde edilen anlamsal gömme modülü ve toplu varlık çözümleme aşamaları detaylandırılmaktadır.

4.1 Doküman Gömme Modülü

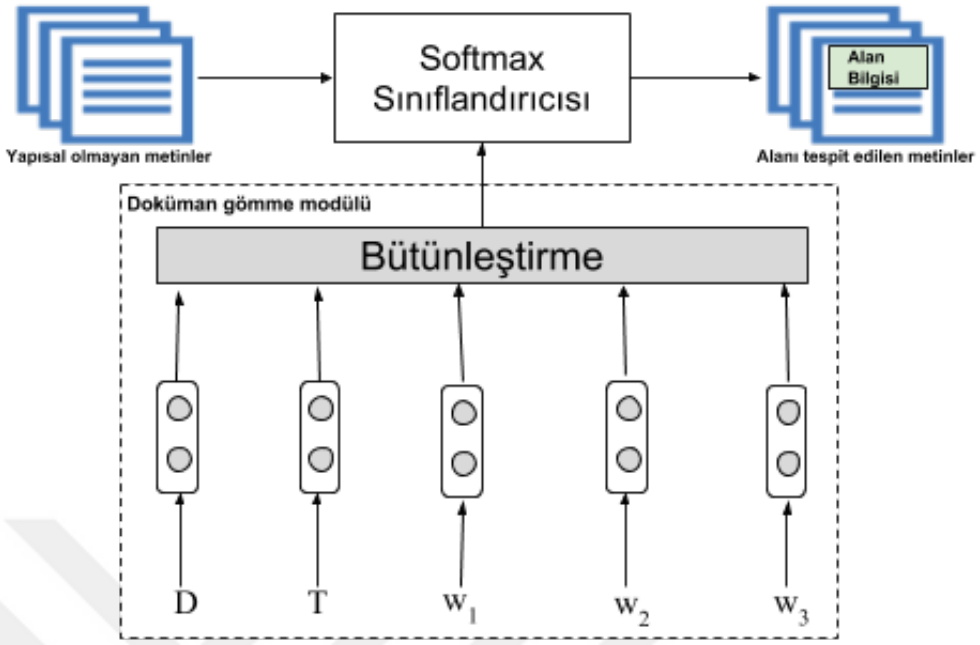
Alan tespiti literatürde doküman sınıflama ve konu tespiti olmak üzere farklı terimler kapsamında günümüze kadar fazlasıyla çalışılmıştır (Dai et al., 2015). Bu bağlamda LDA (Blei et al., 2003) iyi bilinen geleneksel bir yöntem olarak bilgi temsili alanında sıklıkla çalışılmıştır. Bağlam eşlemede kullanılan diğer bir çok popüler çalışmaya örnek vermek gerekirse TF-IDF ağırlıklı Vektör Uzay Modeli (Salton and Buckley, 1988) ,olasılığa dayanan varlık bağlam modeli (Han and Sun, 2011) ve LDA modeline dayanan Tematik Bağlam Mesafesi olarak verilebilir.

Kelime gömme, büyük bir korpus içindeki her kelime için sayısal bir temsildir ve eş anlamlılar, zıt anlamlılar ve analojiler gibi sözcük çiftleri arasındaki ilişkileri yakalayabilmektedir (Mikolov et al., 2013a). Ancak, konu tespiti gibi bazı durumlarda, kelime yerleştirme, belgeler arasındaki benzerliği tahmin etmede kullanışlı olmayabilir. Bu nedenle, Doc2Vec (Le and Mikolov, 2014) modeli, kelime toplamalarının geliştirilmiş bir sürümü olarak sunulmaktadır. Dahası, popüler doküman gömme çalışmaları (Kusner et al., 2015) paragrafları ve dokümanları düşük boyutlu vektör uzayında temsil edebilmektedir. Bu çalışmalar bağlam eşleme ve alan tespiti için kritik çalışmalardır ancak son zamanlardaki bir çalışma Doc2Vec modelinin Varlık Bağlama yöntemi içinde kullanılmasının başarıyı artırdığını göstermektedir (Zwicklbauer et al., 2016b). Temel olarak Doc2Vec modeli her giriş metni için ayırtedici bir paragraf kimliği olan başka bir özellik vektörü eklemektedir.

Bu tez kapsamındaki önerilen yöntem doküman tespiti için başarısı kanıtlanmış Doc2Vec modelini kullanarak yapısal olmayan metinlerin doküman gömme yöntemi ile alanlarını tespit etmektedir. Doküman gömme modeline dayandığı için önerilen bu yöntem alana özgü hazırlanmış bir doküman yığınında kullanılabilir. Ancak bilgi tabanındaki varlıklarla etiketli atıfların arasında kolay eşleşme sunduğu için bu tez kapsamında alana özgü olarak ayrılmış Vikipedi yığınları kullanılmıştır.

Doküman gömme modülünde yer alan Doc2Vec yöntemini oluşturan Paragraf Vektörünün Dağıtılmış Belleği (PV-DM) ve Paragraf Vektörünün Dağıtılmış Çanta (PV-DBOW) olmak üzere iki ana model vardır. PV-DM modeli paragrafın konusunu hatırlar ve TaggedDocument elemanları[23] ile gensim uygulamasını kullanır. Bu nedenle, verilen metnin konusunu belirten bir başka vektörü de mevcut vektör uzayına ekler.

[23] https://radimrehurek.com/gensim/models/doc2vec.html



Şekil 4.2: Alan tespitinin genel yapısı.

Şekil 4.2 alan tespitinin genel yapısını göstermektedir. Kelime gömme modelinden farklı olarak kelime vektörleri $w_1, w_2, w_3, ..., w_N$ ile T metin vektörü birleştirilmiştir. Doküman gömme modelinin girdisi olan her metin için bir paragraf numarası ve alan vektörü D vardır. Bütün vektörlerin birleştirilmesinden sonra öğrenme adımında Softmax sınıflandırıcısı ile verilen metnin alanı tespit edilmektedir. Genel olarak alan tespiti denklemi aşağıda verilmiştir.

$$t_i^* = argmax_i(Softmax(d_j, t_i)) \tag{4.1}$$

Bu denklemde t_i^* girilen metin $t_i, i \in T$ için en yüksek Softmax değeri verilen $d_j, j \in D$ alanı için tespit edilir. Böylece alanı tespit edilmiş metindeki atıflar için diğer alanlardaki aday varlıklar filtrelenmiş olmaktadır. Veri uzayının verimli bir şekilde düşmesi daha etkin bir global çözümleme aşamasının sağlanmaktadır. Sonraki alt bölümde global çözümleme aşamasının girdisini oluşturan anlamsal gömme modeli açıklanacaktır.

4.2 Anlamsal Gömme Modülü

Anlamsal gömme sayesinde varlıklar yada ilişkilerin daha genelleştirilmiş temsilleri düşük boyutlu vektör uzayında saklanabilmektedir (Ji et al., 2015).

Bugüne kadar sıklıkla çalışılan anlamsal gömme modelleri yapay sinir ağları (Bordes et al., 2013), matrisin çarpanlarına ayırma (Chang et al., 2014) ve Bayesian kümeleme (Sutskever et al., 2009) olmak üzere üç kategoride incelenmektedir (Yang et al., 2014). Bu tez kapsamında önerilen yöntemde ana olarak RDF kavramlarına odaklanıldığı için RDF2Vec modeli (Ristoski and Paulheim, 2016) temel alınmıştır.

RDF2Vec modeli Word2Vec (Mikolov et al., 2013a) kelime modelinin RDF izdüşümlerine uyarlanması ile elde edilen bir sinir ağları modelidir. Word2Vec modelindeki esas yaklaşım metin içindeki anlam olarak birbirlerine benzer kelimelerin modelin eğitilmesinden sonra vektörel uzayda daha çok yaklaştığı gözlenmektedir. RDF modelinde ise kelimeler yerine varlıklara ve ilişkilerin sorgulanması ile elde edilen anlamsal izdüşümler eğitilmektedir. Böyle bir yaklaşımın RDF çizge verisine uygulanması için öncelikle çizgenin varlık-ilişki dizilerine dönüştürülmesi gerekmektedir. Böylece bu varlık-ilişki dizilerini, Sinir Ağı Dili modellerinde eğiterek RDF çizgesindeki her bir varlığı Gizil Özellik Uzayı (Latent Feature Space)'ndaki nümerik değerler vektörü şeklinde temsili sağlanmaktadır. RDF2VEC ile anlamsal ilintililiğin ölçülmesi için adımlar aşağıdaki gibi listelenmiştir.

1. RDF çizgelerinin seçilen strateji (Weistfeiler-Lehman Subtree RDF Graph Kernels, graph walks) ile varlık-ilişki dizilerine çevrilmesi
2. Varlık-ilişki dizilerinden Sinir Ağı Dil Modeli (Neural Language Model) yöntemlerinden (CBOW, Skip-Gram) biri ile modelin kurulması
3. Sinir Ağı Dil modelinden iki varlığın anlamsal ilintililiği Softmax fonksiyonu kullanılarak hesaplanması

Bu bölümün sonraki alt bölümlerinde listelenen varlık-ilişki dizi çıkarma, sinir ağı modelinin belirlenmesi ve anlamsal ilintililiğin hesaplanması adımları detaylandırılmaktadır.

4.2.1 Varlık-ilişki dizileri

Anlamsal gömme modelinin oluşturulmasındaki birinci adımda varlık-ilişki dizilerini elde etmek için HDT[24] aracı ile oluşturulmuş bilgi tabanları üzerinde SPARQL sorguları ile her bir varlık için belirli bir derinlikte patikalar çekilmiştir.

[24]http://www.rdfhdt.org/

Örneğin bilgi tabanı yapısı sadece **rdf:type** ve **rdfs:subClassOf** ilişkilerine sahip olduğunda SPARQL sorgusunun çıktısı aşağıdaki şekilde olmaktadır.

```
@prefix rdf:   <http://www.w3.org/1999/02/
                                22-rdf-syntax-ns#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dbo: <http://dbpedia.org/ontology/>.
@prefix dbr: <http://dbpedia.org/resource/>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.

dbr:Citizen_Kane rdf:type dbo:Film . dbo:Film ->
rdfs:subClassOf dbo:Work . dbo:Work ->
rdfs:subClassOf owl:Thing .
```

Şekil 4.3: RDF ontolojisi için varlık-ilişki dizisi örneği

Şekil 4.3 görüldüğü gibi **dbr:Citizen_Kane** varlığı sadece RDF tip hiyerarşisine sahiptir ve elde edilen bu üçlüler dört adet kavramı (dbr:Citizen_Kane, dbo:Film, dbo:Work and owl:Thing) içeren bir dizi oluşturmuştur. Ayrıca "rdf:type" ve "rdfs:subClassOf" ilişkilerine sahiptir. Bu örnekte derinlik üç olarak belirlenmiştir çünkü **owl:Thing** kavramına kadar varlığa ait üç bağlantı vardır. Belirli bir yolu izleyerek dizi oluşturmak yerine çizge çekirdekleri de kullanılabilir (de Vries, 2013). Sonuç olarak, elde edilen bu dizi tek bir satır olarak sinir ağı modeline gönderilmektedir.

4.2.2 Sinir ağı modeli ve anlamsal ilintililik

Varlıklara ait satır satır varlık-ilişki dizileri çıkarıldıktan sonra iki varlığın anlamsal ilintililiği Softmax fonksiyonu kullanılarak iki varlığa ait vektörel temsillerinin koşullu olasılığı olarak hesaplanmaktadır. Bu şekilde ilintililiği 1'e en yakın çıkan varlık çiftinin birbirine daha çok ilintili olduğu anlaşılmaktadır.

$$p(e_0|e_i) = \frac{exp(v_{e0}^{'T} v_{ei})}{\sum_{e=1}^{V} exp(v_e^{'T} v_{ei})} \tag{4.2}$$

Yukarıdaki formül Softmax fonksiyonunun varlık ilintililik hesabı için özelleşmiş hali olarak v_e girdi vektörüne ve $v_e^{'}$ çıktı vektörüne sahiptir. V ise bütün varlıklara ait sözlüğü belirtmektedir. RDF2Vec (Ristoski and Paulheim, 2016) çalışmasında en yüksek başarının elde edildiği durumdaki en uygun ilerleme adımları ve çizge derinliği alınmıştır. Bu tez çalışmasının ilgi alanı RDF2Vec

modelindeki ilerlemeli denemeler ve çizge derinliği olmadığı için bu konudaki detaya yer verilmemiştir.

Varlıklar arasındaki anlamsal ilintililiği belirlemek varlık çözümleme aşamasında aday varlıkların sıralanması adına çok önemlidir ve yukarıdaki Softmax formülü alt bölümlerde yer alan katmanlı mimarilerde kullanılmaktadır.

4.3 Toplu Varlık Çözümleme

Varlık Bağlama yaklaşımları atıf belirlenmesi ve varlık çözümlenmesi alt görevlerini içermektedir. Varlık çözümlemesi aşaması içinde oluşturulan aday varlıkların sıralanması adımını barındırmaktadır.

4.3.1 Aday varlıkların oluşturulması

Atıf etiketli metinlerde yer alan atıflara ait aday varlıkları oluşturmak için DBpedia bilgi tabanını temel alan bir yaklaşım geliştirilmiştir. Bu aday varlıklar için önceden tespit edilmiş atıfları etiketli metinler Vikipedi bilgi kaynağından toplanmıştır. Her bir atıf için, aday varlıkları ve alan bilgileri DBpedia bilgi tabanından sorgulanmıştır. Bunu yapmak için, varlıkların "dct: subject"[25] ve "rdf: type"[26] özellikleri ile birlikte Schema.org[27] alanları kontrol edilmiştir. Daha sonra, alan bilgilerini içeren atıf ve aday varlıklar bir anahtar-değer deposuna kaydedilmiştir.

Vikipedi makaleleri paragraflara ayrılır ve paragrafın herhangi bir açıklamalı varlığı içerdiğini anahtar değer deposunda her paragraf alınmıştır. Belirtilen paragrafta bir veya daha fazla varlık varsa ve bu paragraf ek açıklamalı metin listesinde yoksa, belirtilen paragraf bir belge deposuna yüklenmiştir. Aynı zamanda, paragrafta bulunan her bir söz için Vikipedia anlam karmaşıklığı sayfaları aranmıştır. Herhangi bir söz için herhangi bir anlam karmaşıklığı sayfası varsa, anlam karmaşıklığı olan varlıklar içeren atıf etiketli metinler de saklanmıştır.

[25] http://purl.org/dc/terms/subject
[26] http://www.w3.org/1999/02/22-rdf-syntax-nstype
[27] https://schema.org

4.3.2 UKHA modeli

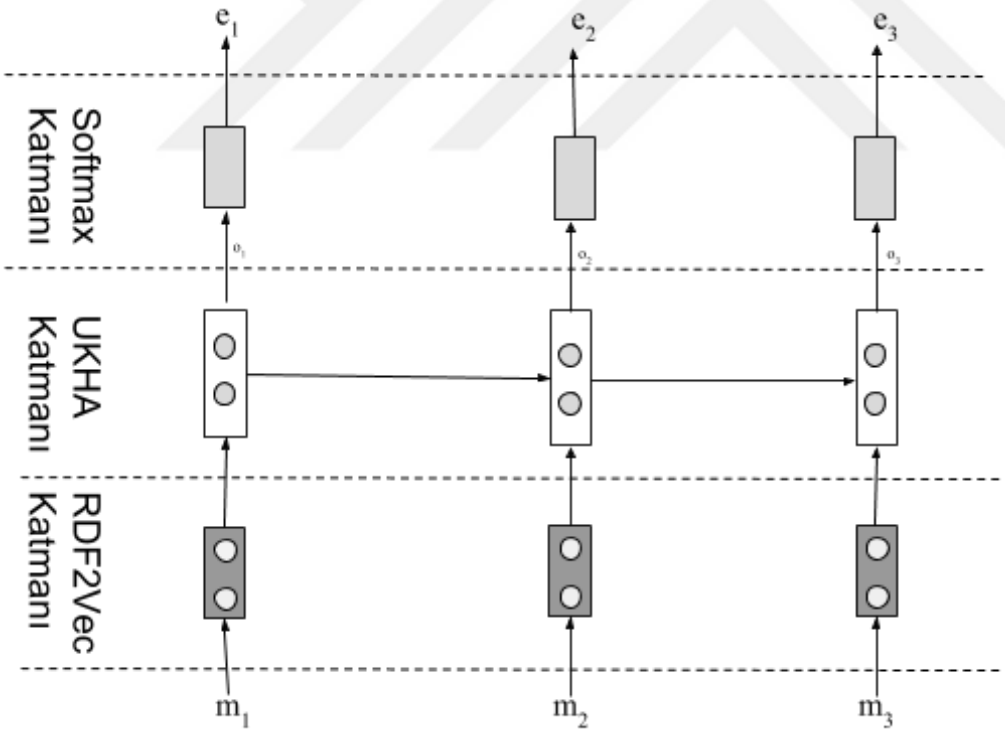
Bu bölümde diziden diziye öğrenme modelinin Varlık Bağlama yöntemine uyarlanmasında belirli bir notasyon kullanılmıştır. Yapısal olmayan bir metin ele alındığında $M = \{m_1, m_2, ..., m_N\}$ atıflar kümesi bu metnin içinde yer almaktadır. Ayrıca bu atıfların karşılığı olarak $E = \{e_1, e_2, ..., e_N\}$ varlıklar kümesini de içermektedir. Burada N sayısı kümedeki toplam sayı yerine o andaki dizideki boyutu göstermektedir. Yani atıf sözlüğü ile varlık sözlüğünün toplam eleman sayısı birbirinin aynısı olmayabilmektedir. Bu tez kapsamında önerilen metodun amacı her bir atıfı onu temsil eden varlık ile $M_i \rightarrow E_i$) eşlemektir. Burada varlıkların içerdiği bilgi tabanı ile atıfların bulunduğu metinlerin alan bağımlı kaynaklar olduğu baştan tespit edilmektedir.

Aday varlıkların sıralanması için Tekrarlayan Yapay Sinir Ağları (TYSA) yöntemlerinden faydalanılmıştır. Tekrarlayan Sinir Ağları tahmin yapmak için girdi bilgilerini diziler halinde belirli bir sırada kullanmaktadır. Bu ağlarda girdi ve çıktı dizileri arasındaki bağımlılık göz önüne alındığından dolayı geleneksel yapay sinir ağı yöntemlerinden ayrılmaktadır. Tekrarlayan Sinir Ağları dizinin her terimi için tekrarlanan görevleri yerine getirerek o ana kadar hesaplanmış olan bilgileri hafızaya almaktadır. Bu modellerin teorik olarak uzun dizi girdileri üzerinde çalışabildiğini gösterse de paratikte başarısız olma eğilimindedir (Bengio et al., 1994). Çok uzun dizi girdileri TYSA modeline verildiğinde en son giren dizi bilgisini hatırlaması tahminlemenin genel başarısını olumsuz etkilemektedir. Bu problemin üstesinden gelinmesi için Uzun Kısa-Vade Hafıza Ağları (UKHA), uzun diziler üzerinde çalışabilen bir bellek hücresi üreterek sadece en son girilen dizi girdilerini değil tahminleme için kritik olan dizi girdilerini dikkate almaktadır (Hochreiter and Schmidhuber, 1997).

Varlıkları çözümlemek için UKHA katmanı atıflardan oluşan girdi ve aday varlıklardan oluşan çıktı dizilerinde çalıştırılmıştır. Her bir atıf için doğru aday varlık tahminlenmesi hesaplarken h_{t-1} gizli durumunu $t-1$ zaman aralığında aşağıdaki denklem ile hesaplanmıştır.

$$
\begin{aligned}
C_t^* &= tanh(W_c[h_{t-1}, m_i^t]) + b_c) \\
\phi_u &= \sigma(W_u[h_{t-1}, m_t]) + b_u) \\
\phi_f &= \sigma(W_f[h_{t-1}, m_t]) + b_f) \\
\phi_o &= \sigma(W_o[h_{t-1}, m_t]) + b_o) \\
C_t &= \phi_u * C_t^* + \phi_f * C_{t-1}
\end{aligned}
\tag{4.3}
$$

Burada C_t^* adayı ile C_t yer değiştirilmesi $tanh$ aktivasyon fonksiyonu Wc üzerinde işletilmiştir. Aynı zamanda, UKHA modelinde ϕ_u güncelleme, ϕ_f ise unutma kapısıyken ϕ_o çıktı kapısıdır. Daha sonra σ eleman bazlı sigmoid fonksiyonu olarak güncelleme, unutma ve çıktı matrislerini hesaplamaktadır. Bunu hesaplarken h_{t-1} gizli durumunu *t-1* zaman aralığında dikkate alarak girdi atıfı m_t üzerinden çalıştırmaktadır.



Şekil 4.4: UKHA modelinin genel yapısı.

Varlık çözümlenmesi için ilk önce UKHA katmanı girdi ve çıktı dizilerinde çalıştırılmıştır. Buradan üretilen h_t gizli durumu *t* zaman aralığında üretilmiştir. Daha sonra aşağıdaki denklemde varlık çözümleme için e_i^* kuralı hesaplanmıştır:

$$e_i^* = argmax_j(logSoftmax(Ohi + b))_j \tag{4.4}$$

Burada e_i^*, bu vektörde en yüksek puanı alan aday varlığı göstermektedir. Yani girdi olarak verilen atıf için en iyi tahminlenmiş aday varlığı E varlık sözlüğünden seçilmiştir. Şekil 4.4, aşağıdakilerden oluşan UKHA katmanlı mimarisinin katmanları aşağıdaki gibi listelenmiştir.

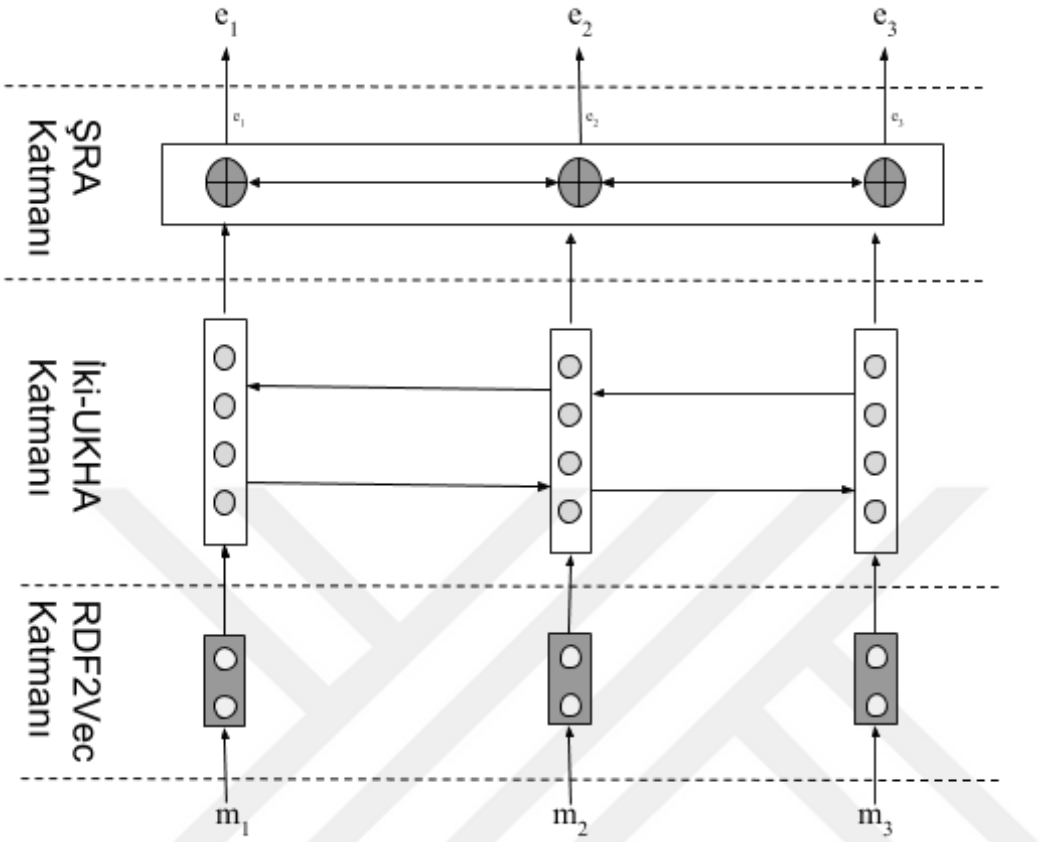
- RDF2Vec katmanı, M atıflar sözlüğündeki m_i atıfını $m_i \in M$ ve bu atıfa karşılık gelen e_i referans varlığını $e_i \in E$ gerçek değerli ve d-boyutlu vektör uzayına aktarmaktadır.
- Gizli h_t ve o_t çıkış vektörlerini üreten bir UKHA modeli *t* ileri geçiş vektörü $\overrightarrow{r_t}$ olarak elde edilmektedir.
- Softmax katmanı ile çıktı vektörü o_t, *t* zaman aralığında *E* varlık sözlüğü üzerinde olasılık dağılımına dönüştürülmektedir.

Girdi olarak kullanılan RDF2Vec katmanı içinde mevcut tez kapsamında ilişkilerin ağırlıkları dikkate alınmamış olur atıf ve varlıkların dizileri üzerinden hesaplamalar yapılmıştır.

4.3.3 iki-UKHA ve ŞRA

Aday varlıkların sıralanması için denenen ikinci modelde iki yönlü UKHA modelinden yararlanılmıştır. UKHA modeli, atıfa ait girdi dizileri için t zaman anındaki diziden başlayarak soldan sağa $\overrightarrow{h_t}$ gizli vektörü ile temsil edilmektedir. Fakat bu kritik bilgileri ters sırayla yani sağdan sola olmak üzere görmezden gelebilmektedir. Hem soldan sağa hem de sağdan sola kritik bilgileri hesaba katmak adına ikinci bir UKHA modeli mevcut modelin tersi yönde çalışacak şekilde eklenmektedir. Daha sonra, bu ileri ve geri UKHA çift modelinden iki yönlü bir UKHA (Graves and Schmidhuber, 2005) modeli oluşturulmaktadır. Böylece iki-UKHA modeli sol ve sağ içeriği ile atıf dizilerini temsil ederrek bu dizilerden daha kapsamlı bilgi toplamaktadır. İkinci bir UKHA modeli mevcut sisteme eklenmesinden dolayı işlem süresi uzasa da bu tez kapsamında doğruluk başarısı daha çok dikkate alınmaktadır. iki-UKHA ve ŞRA modellerinden oluşan melez yapı Şekil 4.5 olarak gösterilmiştir.

- RDF gömme katmanı $m_i \in M$ her bir atıf için ve referans $e_i \in E$ varlığını



Şekil 4.5: iki-UKHA ve ŞRA modellerinden melezlenmiş yöntem.

gerçel değerli bir d boyutlu vektör elde edilmiştir.

- Daha sonra Bi-UKHA modeli gizli vektör h_t için her bir zaman aralığında t dikkate alınmış h_t ile hesap edilerek $S = \{r_1, r_2, ..., r_N\}$ girdi dizisi için ileri $\overrightarrow{r_t}$ ve $\overleftarrow{r_t}$ geri RDF elemanlarından oluşan vektörler oluşturulmuştur.

- Son olarak ŞRA katmanı birbirleriyle bağlı olan model ile anlam karmaşıklığı oluşturan aday varlıklar için en iyi atıf-varlık çiftleri eşleştirilmektedir. Bu adım global bir varlık çözümleme problemi olarak belirli bir alana özgü çözümlenmektedir.

Bu yapıda varlık çözümleme Şartlı Rastgele Alanlar (ŞRA) (Lafferty et al., 2001) kullanılarak yapılmıştır. Bu durumda, iki-UKHA tarafından üretilen özelliklerin üretildiği bir ortamda ŞRA modeli dizi öğrenme modeli olarak kullanılmıştır. ŞRA modelinin hesapladığı logaritmik denklem aşağıda gösterilmiştir:

$$p(e|m) = \frac{exp(lp(m,e))}{\sum_{e'} exp(lp(m,e'))} \tag{4.5}$$

Burada m atıfların girdi dizisini gösterirken e varlıkların çıktı dizileridir. Daha sonra, lp atıf ve varlıkların log potansiyel değerini göstermektedir. Metodun sağlıklı çalışması için potansiyel değerlerin yerel değerleri içermesi gerektirmektedir. Bu yüzden emisyon ve geçiş değerleri Bi-UKHA destekli CRF modelinde uygulanan iki potansiyel değer olarak ele alınmıştır. Daha sonra bu değerlerin log potansiyelleri aşağıdaki şekilde hesaplanmıştır.

$$lp(e,m) = \sum_{i} log\theta_E(e_i \rightarrow m_i) + log\theta_T(e_{i-1} \rightarrow e_i) \tag{4.6}$$

Burada ($log\theta_E$) emisyon değerini ifade etmektedir. Atıfın i indeksinde Bi-UKHA modelindeki gizli durumdan i zaman aralığı için hesaplanmıştır. Daha sonra ($log\theta_T$) geçiş potansiyel değerleri $|E|x|E|$ matrisi P içinde saklanmıştır. Burada E ayırtedici varlıklardan oluşan alana özgü sözlüktür. Bu tez kapsamında PyTorch[28] kütüphanesi kullanılarak UKHA, Bi-UKHA ve ŞRA modelleri gerçekleştirilmiştir. PyTorch dinamik sinir ağı aracı olarak her bir örnek için canlı bir şekilde hesaplanan bir hesap çizgesini sahiptir.

4.3.4 Kodlayıcı-kod çözücü ve dikkat mekanizması

Tez kapsamında aday varlıkların sıralanması için önerilen yöntemin ana yapısı olarak kodlayıcı-kod çözücü sinir ağ modelleri ve dikkat mekanizmasından yararlanılmıştır. Bi-UKHA modelinin daha genel bir sürümü olarak bu model daha sonra Varlık Bağlama probleminin çözümü için kaynak olarak verilen atıf dizilerinin karşılık gelen hedef varlık dizilerinin tahmin edilmektedir. Böylece diziden diziye öğrenme modeli global varlık çözümleme aşamasına dönüştürülerek bir kodlayıcı-kod çözücü mimarisine uyarlanmıştır (Sutskever et al., 2014).

Diziden diziye öğrenme modelleri kodlayıcı ve kod çözücü modelleri olarak iki ana yapıyı içermektedir. İlk model olarak kodlayıcı yapı kaynak olarak atıflardan oluşan giriş dizisini okuyup bunu sabit uzunlukta bir vektöre kodlamaktadır. Daha sonra, ikinci model olan kod çözücü yapısı kodlanan bu vektörü çözerek hedef aday varlıklardan oluşan çıkış dizisini tahmin etmektedir. Tez kapsamında önerilen

[28]http://pytorch.org/tutorials/index.html

yöntemde atıfların dizileri sabit boyutlu vektör gösterimi kodlayıcı modeline beslenmektedir. Sunulan kodlayıcı modelinin (Sutskever et al., 2014) tanımı aşağıdaki denklemde gösterilmiştir.

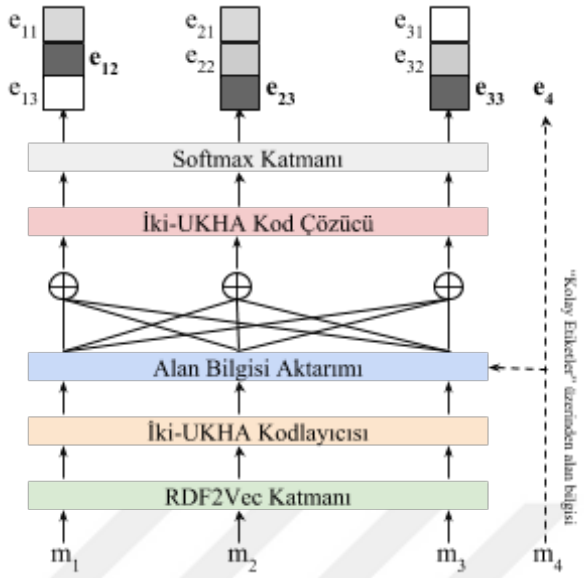
$$\begin{aligned} h_t &= \phi(m_t, h_{t-1}) \\ s_v &= \gamma(\{h1, ..., h_N\}) \end{aligned} \tag{4.7}$$

Burada h_t gizli bir vektör olarak m_t, t zamanında atıf dizisinin vektörüdür ve s_v, gizli durumların dizisinden oluşturulan bir dizi vektördür. Ayrıca ϕ ve γ UKHA modelleri gibi doğrusal olmayan işlevler olarak yer almaktadır. Kod çözücü yapısı koşullu olasılıklara göre varlık sıraları üzerinden $p(e_1, ..., e_N|m_1, ..., m_N)$ dağılımını aşağıdaki denklemde olduğu gibi tanımlamaktadır.

$$p(e_1, ..., e_N|m_1, ..., m_N) = \prod_{t=1} p(e_t|s_v, e_1, ..., e_t - 1) \tag{4.8}$$

Yukarıdaki denklemde N, atıf ve varlık dizilerinin uzunluğudur. Ek olarak s_v tüm girdi atıfları için dizi vektörüdür ve $p(e_t|s_v, e_1, ..., e_t - 1)$ ile temsil edilen koşullu olasılık hesabını göstermektedir. M sözlüğündeki tüm atıflar üzerinde bir softmax fonksiyonu çalıştırılmaktadır. Bu kodlayıcı/kod çözücü modelinde yer alan kodlayıcı fonksiyonunun giriş dizisini sabit uzunlukta bir vektöre kodladığı bir problemle karşılaşabilir ve bu durumda kod çözücü tüm çıkış dizisini tahmin eder. Ayrıca, önceki alt bölümün iki yönlü UKHA modeli, gizli durum h_i kodlanmış bütün girdi dizisinden gelen bilgileri kullanmaktadır. Bununla birlikte, giriş sırasındaki bazı ifadeler, referans varlığın bağlanmasında diğerlerinden daha ayırt edici olabilmektedir. belirli bir zaman adımı. Yapay Makine Çevirisi (Bahdanau et al., 2014) modelindeki vizyonuna alan bilgisini içeren katmanı bir dikkat mekanizması ile besleyen bir yöntem bu tez kapsamında önerilmiştir. Dikkat mekanizması, kod çözücünün çıkış dizisindeki referans varlığını girdi dizisindeki atıfın alan bilgisi kullanarak hangi aday varlığın daha fazla dikkat çekmesi gerektiğini göstermektedir.

Dikkat mekanizmasının kodlayıcı-kod çözücü mimarisi Şekil 4.6 olarak gösterilmiştir. Kodlayıcı modeli sırasıyla RDF2Vec, iki-UKHA ve alan bilgisi aktarılmış dikkat mekanizması katmanlarını içermektedir. Atıf dizisinde geçen m_i atıfı RDF gömme modeline dönüştürülmektedir ve önceki mimari olarak iki-UKHA katmanı üzerinden çalıştırılmaktadır.



Şekil 4.6: Dikkat mekanizmasının kodlayıcı/kod çözücü yöntemi.

Burada c, tüm sıralama dizileri tarafından oluşturulan ve sonraki iki-UKHA modelinin çıktı vektörleri ile birleştirerek aday varlıklardan oluşan diziyi tahmin etmektedir. Şekil 4.6 gösterildiği gibi tez kapsamında önerilen ana varlık çözümleme yöntemi RDF gömme, iki-UKHA modeli ve alan bilgisi eklenmiş dikkat mekanizması olmak üzere üç katmandan oluşan bir kodlayıcı modelden oluşmaktadır. Bu kodlayıcı modeli c bağlam vektörünü üretmektedir. Bundan sonra, bir kod çözücü modeli iki-UKHA ve Softmax katmanları içerir. Buradaki iki-UKHA modeli, bağlam vektörü üzerinde çalışmaktadır ve bu modelin çıktıları, softmax katmanları üzerinde yürütülmektedir.

4.4 Sonuç ve Değerlendirme

Bu bölümde öncelikle alan uyumlu anlamsal ve doküman gömme modellerine dayanan ve herhangi bir bilgi kaynağı eklenerek genişleyebilen ölçeklenebilir bir Varlık Bağlama mimarisi önerilmiştir. Bu mimarinin ana adımları alan tespiti ve global varlık çözümleme olmak üzere iki ana başlıkta incelenmiştir.

Önerilen yöntemin ilk adımı alan tespiti ve global varlık çözümleme aşamalarının girdileri olan doküman ve anlamsal gömme modüllerinin oluşturulmasıdır. Bu modülleri oluşturmak adına önceden eşlenmiş atıf-varlık çiftlerini barındıran atıf etiketli doküman deposu olarak Vikipedi kaynağı kullanılırken bu atıflara referans varlıkları içeren DBpedia bilgi tabanından faydanalınmıştır. Vikipedi kaynağındaki anlamsal kategori sayfalarındaki bilgilerle

DBpedia kaynağında yer alan şema bilgileri örtüştürerek alan bilgilerine göre iki kaynak için ayrı ayrı alan uyumlu veri kaynakları üretilmiştir. Doküman ve anlamsal gömme modellerinin eğitilmesi için gensim ve DL4J kütüphanelerinden yararlanılmıştır.

Üretilen doküman ve anlamsal gömme modelleri sırasıyla alan tespiti ve global varlık çözümleme aşamalarında kullanılmıştır. Alan tespiti için kullanılan yöntem Doc2Vec modeli olmuştur. Birçok alan tespiti yöntemi olmasına rağmen güncel olan Doc2Vec modeli diğer yöntemlere karşı başarı sağladığı için bu tez kapsamında bu model üzerinde durulmuştur. Yapısal olmayan metin bu aşamadan geçtiğinde bağlam olarak hangi alanda olduğu tespit edilmesi ile birlikte bir sonraki aşama olan global varlık çözümleme aşamasına geçilmektedir. Global varlık çözümleme aşamasında alan tespiti yapıldığı için aday varlıkların sayısı azalmış olmaktadır. Bu durum da global varlık çözümleme aşamasını belirli bir alan için yapılmasından dolayı işlem süresini azaltmaktadır.

İleriki çalışmalarda, önerilen mimarinin gerçekleştirim kapsamının genişletilebileceği pek çok yön bulunmaktadır. Bu kapsamda, öncelikli olarak kısa vadede ele alınabilecekler aşağıda listelenmiştir:

- Alan tespiti için kullanılan Doc2Vec modelinin güncel çalışmalarla kıyaslanarak başarısı gözlenecektir. Bu bağlamda geleneksel yöntemler ve güncel yöntemler harmanlanarak daha iyi sonuç verebilecek karma modelin geliştirilebilinir.
- Bilgi tabanı gömme modeli olarak kullanılan RDF2Vec modeli ile güncel çizge gömme modellerinden başarısı kıyaslanabilir. Son yıllarda çok fazla bilgi tabanı gömme modeli geliştirilmesine rağmen bu tez kapsamındaki bilgi tabanı RDF elemanlarında oluştuğu için bu aşamada RDF2Vec modeli tercih edilmiştir. İleri zamanlarda bilgi tabanı gömme modelleri ile RDF2Vec modeli arasında karşılaştırma yapılabilir.
- Anlamsal gömme modellerinde mevcut durumda RDF elemanlarında atıf ve varlık çiftleri özne ve nesne durumları açısından dizilişleri incelenmiştir. Atıf ve varlıklar arasında bilgi tabanındaki ilişkiler mevcut tez kapsamında değinilmemiştir. İleriki zamanlarda ilişki yapısının genel çözümleme yöntemindeki etkisi gözlenebilir.

Önerilen mimarinin en önemli özelliği farklı pek çok bileşeni üzerinden genişletilebilir yapıda olmasıdır. Dolayısıyla, tez çalışmasının gerçekleştiriminin devam edebileceği pek çok alan bulunmaktadır.

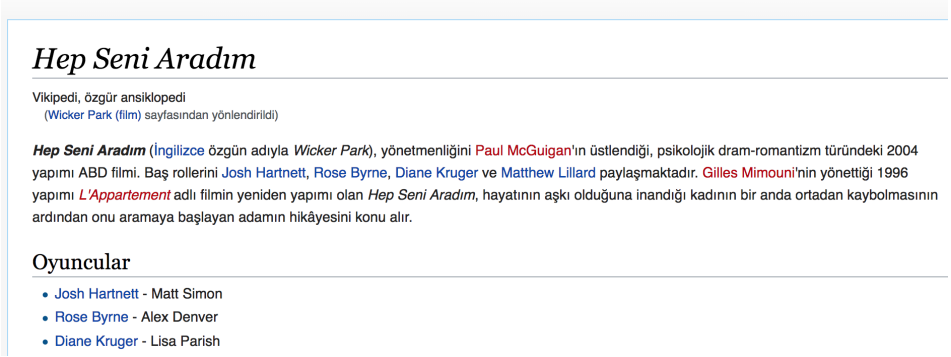
İlk aşamada farklı doküman depoları ile farklı bilgi tabanı kaynakları kullanılarak tez kapsamı dışındaki alanlar için doküman ve bilgi gömme modelleri geliştirilebilir. Bu aşamada doküman yığınındaki atıflar ile bilgi tabanındaki bu atıflara karşılığı gelen varlıkların öncelikle eşlenmesi gerekebilir. Daha sonra bu gömme modelleri bu tez kapsamında önerilen ölçeklenebilir mimariyle bütünleştirilerek farklı doküman ve bilgi tabanı modellerinin genel başarıya etkisi gözlenebilir.

Son olarak değerlendirme veri kümesi oluşturan araç olan WeDGeM aracının sonradan adapte edilen alan doğrultusunda genişletilerek o alana özgü de değerlendirme kümesi oluşturulabilir. Önerilen yöntemin diğer en son teknolojik sistemlerle karşılaştırabilmesi için değerlendirme kümesinin Gerbil aracına uygun veriler biçiminde üretilmesi gerekmektedir.

5 ÖNERİLEN YÖNTEMİN DEĞERLENDİRİLMESİ ve TARTIŞMA

Bu bölümde, gerçekleştirilen mimari olan GnoSSEA yönteminin başarımının değerlendirilmesi için öncelikle üretilen değerlendirme veri kümeleri tanıtılacaktır. Daha sonra ilk aşamada tek bir alan için diğer bilinen sistemlerle karşılaştırılma yapılacaktır. En son olarak farklı alanlar için genişletilerek oluşturulan değerlendirme kümelerindeki başarı diğer sistemlerle karşılaştırılmaktadır.

Değerlendirme adımları hem eğitim hem de test kümelerinin üretimi ile başlar. Bu kümelerin üretilmesinin ardındaki sezgi, alanı tespit edilen metinlerdeki atıflar ile verilen bilgi tabanlarındaki aday varlıklar arasındaki bir eşleşmenin olduğu kabul edilmektedir. Bunu yapmak için, DBpedia'yı genel amaçlı doküman yığını olarak seçilen Vikipedi ile uyumlu bir bilgi tabanı olarak seçilmiştir. Bu bilgi kaynakları zaten birbirine bağlı olduğundan, DBpedia'dan alan odaklı bilgi tabanı çekilip Vikipedi tespit edilen belgeleri ile atıfları etiketlenmiştir. Alana uyumlu bilgi tabanları elde edebilmek için popüler bilgi tabanlarında ortak olarak kullanılan Schema.org[29] kaynağından yararlanılmıştır. Bu şekilde DBpedia bilgi tabanından çıkarılan alana özgü bilgi tabanları Schema.org hiyerarşisini kullanılarak elde edilmiştir. Daha sonra alana özgü elde edilen bu varlıklara ait atıfları etiketli metinler Vikipedi sayfalarından çekilmiştir.



Hep Seni Aradım

Vikipedi, özgür ansiklopedi
(Wicker Park (film) sayfasından yönlendirildi)

Hep Seni Aradım (İngilizce özgün adıyla *Wicker Park*), yönetmenliğini Paul McGuigan'ın üstlendiği, psikolojik dram-romantizm türündeki 2004 yapımı ABD filmi. Baş rollerini Josh Hartnett, Rose Byrne, Diane Kruger ve Matthew Lillard paylaşmaktadır. Gilles Mimouni'nin yönettiği 1996 yapımı *L'Appartement* adlı filmin yeniden yapımı olan *Hep Seni Aradım*, hayatının aşkı olduğuna inandığı kadının bir anda ortadan kaybolmasının ardından onu aramaya başlayan adamın hikâyesini konu alır.

Oyuncular

- Josh Hartnett - Matt Simon
- Rose Byrne - Alex Denver
- Diane Kruger - Lisa Parish

Şekil 5.1: Vikipedi varlık sayfası örneği

Şekil 5.1 gösterilen örnek *Wicker Park* atıfı Vikipedi sayfasındaki örnek metinde geçmektedir. Bu cümlede geçen atıfın *Wicker_Park_(film)*[30] varlığı ile eşleşmesi beklenmektedir.

1. "Wicker Park is a film directed by Paul McGuigan and starring Josh Hartnett"

[29] https://schema.org/
[30] http://dbpedia.org/resource/Wicker_Park_(film)

2. $Wicker\_Park\_(film) Paul\_McGuigan\_(filmmaker) Josh\_Hartnett$

3. $dbr : Wicker\_Park\_(film) dbr : Paul\_McGuigan\_(filmmaker) dbr : Josh\_Hartnett$

Yukarıda gösterilen üç metin dizisi eğitim aşamasında kullanılacak girdi örneklerini göstermektedir. Birinci örnek metin Vikipedi kaynağından alana özgü atıf için çekilmiş ve doküman gömme modelinde kullanılmıştır. İkinci terim dizisi ilk örnekteki atıfın geçtiği metinde etiketlenen atıfların dizisidir. Son olarak üçüncü dizi örneği de ikinci atıf dizisinin karşılığı olan varlıkların verilen bilgi tabanından sorgulanması ile elde edilmiştir.

Şekil 5.2 Öğrenme ve değerlendirme aşamalarının gerçekliğini artırmak için atıflara ait olan Vikipedi anlam ayrımı sayfalarından yararlanılmıştır.

Şekil 5.2: Vikipedi anlam ayrımı sayfa örneği

Örneğin, *Wicker Park* atıfı Vikipedi anlam karmaşıklığı sayfasında *Wicker_Park_(film)*[31], *Wicker_Park_(Chicago_park)*[32] ve *Wicker_Park_(soundtrack)*[33] olmak üzere üç farklı alana ait varlık ile ilişkilendirilmiştir. Ayrıca diğer atıf olan *Paul_McGuigan* iki farklı aday varlığa sahiptir. Verilen örnek cümlede sadece son atıf olan *Josh_Hartnett* bir aday varlığa sahiptir. Bu duruma tez kapsamında "kolay etiket" adı verilmiştir. Son atıfın kolay etiketli olması sayesinde bu metnin kolaylıkla sinema alanına özgü olduğu makine tarafından anlaşılabilmektedir.

[31] http://dbpedia.org/resource/Wicker_Park_(film)
[32] http://dbpedia.org/resource/Wicker_Park_(Chicago_park)
[33] http://dbpedia.org/resource/Wicker_Park_(soundtrack)

Sonraki aşamalarda sadece sinema alanı için oluşturulan değerlendirme kümesi ile üç farklı alan için türetilmiş anlam karmaşıklığı düşük ve yoğun olmak üzere iki farklı değerlendirme kümesi de tanıtılacaktır. Değerlendirme kümelerinin tanıltılmasından sonra elde edilen bu üç küme için son teknoloji Varlık Bağlama sistemleri ile bu tez kapsamında önerilen yöntem karşılaştırılacaktır.

5.1 Değerlendirme Veri Kümeleri

Varlık Bağlama çalışmalarında MSNBC (Cucerzan, 2007b), IITB (Kulkarni et al., 2009) ve Wikilinks (Singh et al., 2012) gibi birçok genel amaçlı değerlendirme veri kümeleri bulunmaktadır. MSNBC ve IITB kümeleri elle etiketlenmiş alan bağımsız popüler web dokümanlarını içermektedir. Elle etiketlenmiş kümelerdeki problem belli başlı varlık türlerini etiketleme ve çok az sayıda ortak karara göre işaretleme yapılması olarak gösterilebilinir. Bu eksikliklerin giderilmesi için Wikilinks çalışması Vikipedi kaynağından geniş ölçekli otomatik etiketlenmiş veri kümesi sunmaktadır. Ayrıca Wikilinks etiketli varlıklara ait insan, organizasyon ve yer özelliklerini de içermektedir.

Navigli (Navigli, 2013) çalışmasında alana özgü bilgi kaynaklarının Varlık Bağlama yöntemlerinin daha etkin performansa ulaşabilmesi için gerekliliğini ve öne çıkan bir süreç olduğunu vurgulamıştır. Ancak bu aşamada yöntemlerin farklı dillere ve farklı alanlara özgü bilgi tabanlarının eksikliğinden dolayı karşılaştırılabilecek bir ortam yada veri kümesi bulunmamaktadır. Bu bölümde bu eksikliğin üstesinden gelebilmek için geliştirilen WeDGeM (Inan and Dikenelli, 2017) aracı alana ve dile özgü değerlendirme veri kümesi üretilmiştir. WeDGeM Vikipedi anlam ayrımı sayfalarını kullanarak yeterli anlam karmaşıklığını (Strassel et al., 2008) sağlamak ve varlık etiketleme araçlarını adil ve nesnel bir şekilde karşılaştırılması için olanak tanımaktadır.

WeDGeM aracı ile oluşturulan sinema alanına özgü etiketlenmiş veri kümesi Türkçe ve İngilizce dilleri için hazırlanmıştır. Hızle basit bir şekilde hazırlanan bu veri kümesi daha sonra çok bilinen varlık etiketleme yaklaşımları için değerlendirme çerçevesi olan GERBIL aracına monte edilerek Türkçe dil desteği veren Babelfy (Moro et al., 2014a) ve DBpedia Spotliht (Mendes et al., 2011) olmak üzere iki yaklaşımda denenmiştir. Bu denemenin amacı seçilen yaklaşımları karşılaştırmak yerine oluşturulan veri kümesinin uygulanabilirliğini göstermektir.

Algoritma 1 görüldüğü gibi Vikipedi makaleleri W_a Vikipedi yığınlarından[34]

[34]https://dumps.wikimedia.org/

Veri: W_a Vikipedi makalesi, A_t etiketli metin, A_d etiketli anlam ayrımı metni, D_e varlık sözlüğü, L_e varlık listesi, p paragraf, l dil, c kategori
Sonuç: domain-specific annotated text including entities
$W_a \leftarrow setLangCat(l, c)$;
$D_e \leftarrow generateEntDict(c)$;
for ***each*** *p in* W_a **do**
  if *hasEntity(p,* D_e*) and IsDistinct(p)* **then**
    $L_e \leftarrow extractEntity(p, D_e)$;
    $A_t \leftarrow generateText(p, L_e)$;
    for ***each*** *e in* L_e **do**
      if *hasDisambiguationPage(e)* **then**
        $A_d \leftarrow createAmbiguousText(p, e)$;
      else
        anlam karmaşıklığı sayfası değil
      end
    end
    $A_t \leftarrow adjustAmbiguity(A_d)$;
  else
    uygun bir sayfa değil
  end
end

Algoritma 1: Değerlendirme veri kümesi oluşturan algoritma

seçilen her bil doğal dil l Vikipedi kategori sayfası c için elde edilir. Bu metinlerdeki bilgi kutularından DBpedia ile doğrulanarak varlık sözlüğü D_e verilen alana özgü üretilir. Üretilen bu varlık sözlüğü indekslenip her bir Vikipedi makalesindeki etiketli paragraflarda sorgulanır. Alana özgü varlıklara ait etiketlenmiş ve daha önce listede yer almayan metinler A_t etiketli metin listesine eklenir. Mevcut metinlerdeki varlıklar L_e varlık listesine eklenerek algoritmada vikipedi anlam ayrımı sayfalarında aranır.

Vikipedi anlam ayrımı sayfaları da normal sayfalara uygulanıldığı gibi paragraflara ayrılarak yine etiketli metinlerde varlık sözlüğündeki varlıklar varlığı sorgulanır ancak algoritmanın sadeliği açısından aynı işlemler anlam karmaşıklığı içeren sayfalar için ayrıca yazılmammıştır. Eğer anlam karmaşıklığı içeren sayfa varsa bu metin bu sefer A_d etiketli anlam ayrımı metin listesinde tutulur. Son adımda belirlenen anlam karmaşıklığı oranına göre iki listedeki metinler harmanlanarak istenilen karmaşıklık elde edilir.

Algoritmadan elde edilen veri kümeleri bağlantıda[35] yer almaktadır ve GERBIL[36] değerlendirme çerçevesinde istenilen varlık etiketleme yöntemi için kullanıma hazır biçimde bulunmaktadır.

[35]https://github.com/einan/eval4J
[36]http://aksw.org/Projects/GERBIL.html

Çizelge 5.1: Değerlendirme kümelerinin özellikleri.

Alan	**Dil**	#**Belge**	#**Varlık**
Sinema	EN	945	3648
Sinema	TR	824	3182

Bu alt bölümde sadece sinema alanında ve farklı alanların birleştirilmesinden türetilmiş değerlendirme veri kümeleri detaylandırılacaktır. Daha sonra ilk aşamada tek bir alan için diğer bilinen sistemlerle karşılaştırılma yapılacaktır. En son olarak farklı alanlar için genişletilerek oluşturulan değerlendirme kümelerindeki başarı diğer sistemlerle karşılaştırılmaktadır.

5.1.1 Sinema alanına özgü değerlendirme kümesi

Sinema alanında değerlendirme kümesi oluşturulurken öncelikle Türkçe[37] ve İngilizce[38] dillerine ait Vikipedi yığınları etiketli metin oluşturmak için indirilmiştir. Daha sonra etiketli metinlerin elde edilme süreci sinema alanında başlatılarak İngilizce ve Türkçe filmler kategorilerine ait metinler Vikipedi yığınlarından ayrıştırılmıştır.

Çizelge 5.1 görüldüğü gibi dil, etiketli belge sayısı ve varlık sayıları sinema alanı için gösterilmektedir. Bilgi tabanı bağımsız Varlık Bağlama sistemlerinde Türkçe desteği olmadığı için bu tez kapsamında İngilizce veri kümesi ele alınıp bu veri kümesinde 945 etiketli doküman bulunmaktadır. Vikipedi bilgi kutularından çekilerek kolay olması açısından DBpedia destekli sinema alanına özgü bilgi tabanındaki eşleştirilmiş ilgili varlıklar film isimleri, yönetmenler ve oyuncular ile ilgilidir. Bu varlıklara ait varsa anlam ayrımı sayfalarından üretilmiş etiketli metinle müzik ve lokasyon gibi farklı alanlarda bulunmaktadır. Vikipedi yığınından etiketli metnin oluşturulmasına geçen bütün adımları kapsayan ortalama süreler de saniye cinsinden verilmiştir. Örneğin, metinler için indirilmiş Vikipedi yığınından etiketli metin çıkarımı her iki dilde ortalama her belge için 0.431 saniye sürmektedir.

Anlam karmaşıklığı oranı, anlam karmaşıklığı olan her bir varlık için bütün Vikipedi anlam ayrımı sayfalarının bütün belgelere bölünmesi ile elde edilmiştir. Örnekte görüldüğü gibi *Wicker Park* atıfı sinema alanında *WickerPark_(film)* varlığı ile eşleşirken 2 alternatif aday varlığa sahiptir. Mevcut durumda İngilizce metinlerin anlam karmaşıklık seviyesi yüzde 28.51 olarak belirlenmiştir. Ellis ve arkadaşlarının çalışması (Ellis et al., 2013) yüzde 13 karmaşıklık gösterirken bu tez

[37] https://dumps.wikimedia.org/trwiki/20170420/
[38] https://dumps.wikimedia.org/enwiki/20170420/

kapsamında elde edilen değer daha yüksektir.

5.1.2 Farklı alanlar için genişletilmiş değerlendirme kümesi

Elle etiketlenen metinler yanlı olma eğilimindedir, çünkü insanlar genellikle varlık açıklamaları için bilinen terimleri seçer. Ayrıca, bu açıklama süreci, popüler olmayan terimler için bazen gürültülüdür. Bu nedenle, Wikipedia, çoklu kullanıcının küratörlüğünü yaptığı ve yapılandırılmış bir ek açıklama sürecini içerdiği için seçilmelidir. MSNBC (Cucerzan, 2007b), IITB (Kulkarni et al., 2009) ve Wikilinks (Singh et al., 2012), genel varlık ek açıklama görevleri için deneysel veri kümeleri önerir. Wikilinks, Vikipedi'ye bağlantılar yoluyla otomatik olarak oluşturulmuş büyük ölçekli bir etiketli corpus sağlar. Wikilinks, çok büyük bir koleksiyonu tanımlamak için otomatik bir yöntem sunmaktadır.

Anlam karmaşıklığı, belirsiz ve benzersiz varlıklar arasındaki orandır ve varlık etiketleme araçları için daha gerçekçi bir ortam sağlar li2012linguistic. Belirsizliği ayarlamak ve belirli alanlar için açıklamalı metinler oluşturmak için, belirli alan adları için en son Wikipedia dökümü İngilizce'nin[39] çıkarılan (Inan and Dikenelli, 2017) adlı yeni bir çalışma bu tez kapsamında geliştirilmiştir. Bunu yapmak için Wikipedia kategori sayfalarını ve Schema.org "schema: CreativeWork"[40] sınıfını kullanılmıştır. Ayrıca, seçilen alan adlarında Vikipedi ayırma sayfalarının kullanıldığı belirsiz bir ortam sağlarlar. Örnek olarak, *Wicker Park* 'un bir Wikipedia belirsizleştirme sayfası [41] vardır ve filmdeki belirsizliği artırmak için kullanılabilir. Açık alanlı bir bilgi tabanından alana özel bir vektör temsili oluşturmak için, her alan için bilgi tabanının yapısını tanımlamak zordur. Örneğin, kitap, film ve müzik alan adlarının farklı özellikleri vardır ve alanın rahatlığı için her bir mülkün araştırılması gerekir. Belirtilen bilgi tabanından bağımsız bir özet öğesi başlatarak alana özgü semantik ve belge yerleşimleri oluşturmak için genel bir yöntem önerilmektedir.

Çizelge 5.2, üç alan için eğitim veri kümelerini göstermektedir. Bu alan adlarının belirsizlikleri, Wikipedia'ya ilişkin sayfalara göre düşük veya yüksek olarak iki türlü ayarlanmıştır. Değerlendirme veri setinin belirsizlik oranı, tüm anlam karmaşıklığı olan varlıkların, söz konusu alan için çıkartılan toplam tekil varlık sayısına bölünmesi olarak hesaplanmıştır. Düşük belirsiz veri kümeleri

[39] https://dumps.wikimedia.org/enwiki/20170420/
[40] https://schema.org/CreativeWork
[41] https://en.wikipedia.org/wiki/Wicker_Park

Çizelge 5.2: Öğrenme veri kümelerinin özellikleri.

Alan	**Anlam Karmaşıklığı**	**#Doküman**	**#Varlık**
Kitap	Düşük	15620	40584
Kitap	Yüksek	14232	61454
Sinema	Düşük	25433	54890
Sinema	Yüksek	22121	127221
Müzik	Düşük	21028	63686
Müzik	Yüksek	19555	101898

aday varlıkların yüzde 25-30'unu belirsizleştirme sayfalarını içeriyor olsa da, yüksek anlam karmaşıklığına sahip olan veri kümeleri adayların yüzde 75-80'ini içermektedir. Düşük ve yüksek belirsizlik düzeyleri için iki değerlendirme veri seti hazırlanır ve her veri kümesi alan başına düşen belgelerin yüzde 33'ünden oluşur. Bu nedenle, Varlık Bağlama sistemlerini değerlendirmek için daha gerçekçi belirsiz bir veri kümesi oluşturulabilir.

Tabloda görüldüğü gibi her alan için öğren veri kümesi oluşturulurken uygulanan adımları sinema alanı için şu şekilde özetlenebilir. Sinema alanındaki filmler, yönetmenler ve başrol oyuncuları, Wikipedia makalelerinin bilgi kutularından çıkarılır ve DBpedia tarafından referans varlıklar ile eşleştirilir. Bu varlıkların açıklayıcı sayfaları, film alanı için değerlendirme veri kümesindeki belirsizlik oranını artırmak için müzik ve kitaplar dahil olmak üzere diğer alanlarda çıkarılır.

5.2 Değerlendirme Sonuçları

Değerlendirme veri kümelerinin detaylandırılmasından sonra bu kümeler üzerinde son teknolojik araçların bu tez kapsamında önerilen yöntem ile karşılaştırılması yapılmıştır. Son teknoloji Varlık Bağlama sistemleri ile önerilen yöntemin karşılaştırılması için GERBIL aracı kullanılmıştır. GERBIL, çok sayıda veri kümesi ve tek tip ölçüm yaklaşımları kullanarak Varlık Bağlama sistemlerinin çevik karşılaştırması için kullanımı kolay bir web tabanlı platform sunmaktadır. GERBIL'a bir araç eklemek için, tüm son kullanıcının, belirli bir gereksinime uyan bir REST arayüzünden web servisi olarak kendi sistemine bir URL vermesi gerekir. Aracın kullanıcı tarafından belirlenen veri kümelerine karşı entegrasyonu ve kıyaslaması, GERBIL platformu tarafından otomatik olarak gerçekleştirilir. Şu anda, GERBIL platformu 9 son teknoloji sisteme ve 11 veri kümesine sahip olarak bu araçların değerlendirmesine olanak sağlamaktadır. Doğal Dil Programlama Değişim Biçimini (NIF) temel alan GERBIL aracına NIF biçimdeki veri kümeleri ve etiketleme araçları için Java sınıfları sağlamaktadır. Bu tez kapsamında önerilen

yöntemin web servisi olarak GERBIL aracına bağlanması yerine sunulan JAVA sınıfları ile değerlendirme süreci gerçekleştirilmiştir.

Bu tez kapsamında karşılaştırılması yapılan araçların detaylandırılması aşağıdaki şekilde yapılmıştır.

- AGDISTIS (Usbeck et al., 2014), aday formları yüzey formlarından tespit edilen ifadeler için seçer ve bu adaylar için bir belirsizlik grafiği oluşturur. Oluşturulan anlaşmazlık grafiği, grafik tabanlı HITS algoritmasında, anlam belirleme adımındaki en iyi söz varlığı çiftlerini eşleştirmek için kullanılır.
- AIDA (Hoffart et al., 2011b), YAGO (Hoffart et al., 2011a) bilgi tabanı üzerinde çalışan aday varlıklar ve yoğun alt algoritmalar arasındaki global tutarlılık hesaplamasına dayanır.
- Babelfy (Moro et al., 2014a) çizge tabanlı bir disambiguation algoritması kullanır ve verilen sözler için aday varlıklar tarafından çevrelenen en yoğun altgrafı bulur. Daha sonra, Babelfy en iyi notu ve varlık çiftini eşleştirmek için en yoğun alt notu kullanır.
- DBpedia Spotlight (Mendes et al., 2011), çok boyutlu bir sözcük boşluğunun varlık başına bir temsile sahip olduğu DBpedia varlık olaylarını içeren bir Vektör Uzay Modelini (VSM) kullanır. DBpedia spot ışığının görmezden gelme görevi, Ters Dönem Frekansı'nı (ITF), terimlerden ziyade aday varlıklara bağlı olan ve VSM'deki sözcüklerle ilişkili aday varlıkların ters orantısına sahip olan bir Ters Aday Frekansı (ICF) haline dönüştürür.
- KEA (Waitelonis and Sack, 2016) sözlük ve bilgi tabanlı bir kombinasyon önererek varlıkların Vikipedi sayfalarındaki birlikte geçiş sıklıklarını analiz eder ve bu sıklıkları Vikipedi bağ yapısı ve DBpedia üzerinde bir çizge yöntemi ile birleştirmektedir.
- PBOH (Ganea et al., 2016), hafifletilmiş Vikipedi istatistiklerine dayanan bir toplu varlık çözümleme sistemi olarak varlıkların birlikte geçiş sıklıklarına hesaplayarak olasılıksal bir çizge modelinde en uyumlu atıf-varlık çiftlerine eşleştirmektedir.
- WAT (Piccinno and Ferragina, 2014) sistemi, TagMe (Ferragina and Scaiella, 2010)'in karmaşık bir sürümüdür. WAT, grafik tabanlı algoritmaya ve oylamaya dayalı bir algoritmadan en iyi söz varlığı çiftinin seçimine bağlıdır.

Değerlendime aracının içinde barındırdığı dokuz Varlık Bağlama sisteminin hepsi ile karşılaştırılma yapılamamıştır. Örnek olarak DoSeR (Zwicklbauer

et al., 2016a) sisteminden sonuçlar alınamadığı için değerlendirme tablosundan çıkarılmıştır. Geriye kalan araçlar da Varlık Bağlama yerine Varlık Tanımlama görevine özelleştiği için verimli bir karşılaştırma olamayacağı için değerlendirmeye alınmamıştır.

5.2.1 Sinema alanında dizi öğrenme sonuçları

Değerlendirme esnasında kullanılan sistemlerin tanıtılmasından sonra kullanılan ölçüm türü olarak GERBIL aracının önerdiği mikro (Mi) ve makro (Ma) metriklerden yararlanılmıştır. Ayrıca GERBIL aracının değerlendirme sürecinde kullanılan D2KB biçiminden yani Bilgi Tabanı için anlam çözümleme olarak isimlendirilen yöntem ile sonuçlar elde edilmiştir. Bu yöntemde metinlerin baştan atıflarla etiketli olduğu varsayılmaktadır. Bu değerlendirmedeki sistemlerin ana amacı atıf etiketli metinlerdeki atıfların verilen bilgi tabanında hangi varlıklarla doğru eşleştiğinin tespit edilmesidir.

Çizelge 5.3: Sinema alanındaki değerlendirme sonuçları.

VB Sistemi	**Mi-F1**	**Mi-P**	**Mi-R**	**Ma-F1**	**Ma-P**	**Ma-R**
AGDISTIS	0.2063	0.2097	0.2031	0.3093	0.3098	0.309
AIDA	0.1485	0.1559	0.1417	0.1975	0.2035	0.1932
Babelfy	0.2101	0.2273	0.1953	0.284	0.2887	0.2812
Dbpedia Spotlight	0.1515	0.1515	0.1515	0.2044	0.2044	0.2044
Kea	0.1478	0.149	0.1466	0.1942	0.1976	0.1922
PBOH	0.2193	0.25	0.1953	0.282	0.282	0.282
WAT	0.2174	0.2451	0.1953	0.3124	0.3439	0.2967
UKHA	0.336	0.342	0.33	0.436	0.45	0.422
iki-UKHA+ŞRA	0.446	0.488	0.41	0.546	0.564	0.53

Çizelge 5.3 üzerinde gösterildiği gibi doğruluk ile Varlık Bağlama görevinin, hassasiyet, anma ve F1 ölçüm değerlerin göre oluşturulan değerlendirmede ölçülen toplam puanlarını göstermektedir. Oluşturulan değerlendirme veri setinin yüksek muğlaklığı nedeniyle tüm puanlar düşüktür. F1 puanlarından görüleceği gibi bu tez kapsamında önerilen yöntemin sadece sinema alanında oluşturulan değerlendirme veri setinde Bi-UKHA + CRF modelini kullanarak en gelişmiş çalışmaları geride bıraktığını göstermektedir.

5.2.2 Farklı alanlar için dizi öğrenme sonuçları

Sinema alanındaki değerlendirme kümesi daha sonra Schema.org hiyerarşisi kullanılarak kitap ve müzik alanları için genişletilerek anlam karmaşıklığına göre

düşük ve yüksek olmak üzere iki farklı değerlendirme veri kümesi oluşturulmuştur.

Yukarıdaki sinema alanı değerlendirme kümesinde denenen sistemler aynı şekilde bu iki farklı veri kümesinde de yine GERBIL aracındaki son teknoloji sistemler ile karşılaştırılmıştır. F1 ölçüsü Makro ve Mikro ölçütlere genelleştirilebilir. Makro ölçütler, tüm açıklamalı dokümanlardaki her bir doküman üzerinde karşılık gelen önemin ortalaması iken, Mikro ölçüler tüm etiketleri birlikte ele alır ve böylece daha fazla etikete sahip belgelere daha fazla önem verir.

Çizelge 5.4: Yüksek anlam karmaşıklığı değerlendirilme sonuçları.

VB Sistemi	**Mic-F1**	**Mic-P**	**Mic-R**	**Mac-F1**	**Mac-P**	**Mac-R**
AGDISTIS	0.348	0.367	0.331	0.483	0.498	0.469
AIDA	0.231	0.249	0.215	0.318	0.325	0.312
Babelfy	0.392	0.396	0.388	0.422	0.432	0.412
Dbpedia Spotlight	0.315	0.315	0.315	0.42	0.422	0.418
Kea	0.215	0.214	0.216	0.291	0.296	0.287
PBOH	0.427	0.425	0.43	0.479	0.486	0.473
WAT	0.406	0.411	0.402	0.433	0.439	0.427
NeuPL	0.502	0.506	0.498	0.539	0.546	0.533
UKHA	0.32	0.326	0.314	0.414	0.422	0.406
iki-UKHA+ŞRA	0.424	0.438	0.41	0.477	0.496	0.46
GnoSSEA	0.586	0.591	0.582	0.55	0.552	0.548

Çizelge 5.4 gösterildiği gibi Varlık Bağlama görevinin hassasiyet, anma ve F1 ölçümlerine göre oluşturulan değerlendirmede ölçülen toplam puanlarını göstermektedir. Oluşturulan değerlendirme veri setinin yüksek belirsizliği nedeniyle tüm puanlar düşüktür. F1 değerleri bu tez kapsamında sunulan yönteme ait bi-UKHA ve etki alanı enjekte edilen dikkat mekanizmasını kullanarak en son teknolojiye sahip çalışmaların, üç alan dahil olmak üzere, oluşturulan yüksek belirsiz veri kümesi üzerinde daha üstün performans gösterdiğini göstermektedir.

Katmanlı yapı açısından ilerlemeli bir şekilde tez kapsamında önerilen GnoSSEA sadece UKHA modelinden oluştuğu durumda güncel sinir ağı modellerinden ve çok bilinen çalışmalardan daha az performans göstermektedir. Bu durum, atıf ve varlık dizilerinin sadece tek yöndeki vektör uzayı temsilleri dikkate alınmasından ve çözümleme için daha kapsamlı bir yöntem kullanılmamasından kaynaklanmaktadır. Varlık çözümleme için ŞRA yöntemi kullanıldığı ve iki yönlü olarak vektör temsillerinin oluşturulduğu iki-UKHA+ŞRA aşamasında popüler yöntemlerden daha başarılı sonuç üretilmektedir. Ancak yine güncel NeuPL çalışmasından daha yüksek performans gösterilememektedir. Son aşamada alan bilgisi dahil edilmiş dikkat mekanizması kullanılarak NeuPL çalışmasından daha iyi sonuç sağlanmaktadır.

Çizelge 5.5: Düşük anlam karmaşıklığı değerlendirilme sonuçları.

VB Sistemi	**Mic-F1**	**Mic-P**	**Mic-R**	**Mac-F1**	**Mac-P**	**Mac-R**
AGDISTIS	0.64	0.67	0.621	0.683	0.688	0.679
AIDA	0.537	0.559	0.516	0.563	0.565	0.562
Babelfy	0.763	0.773	0.753	0.782	0.784	0.781
Dbpedia Spotlight	0.505	0.53	0.482	0.56	0.564	0.556
Kea	0.482	0.49	0.475	0.472	0.476	0.468
PBOH	0.784	0.777	0.792	0.817	0.82	0.815
WAT	0.768	0.772	0.765	0.805	0.809	0.801
NeuPL	0.817	0.82	0.814	0.839	0.84	0.838
UKHA	0.614	0.626	0.603	0.646	0.66	0.632
iki-UKHA+ŞRA	0.76	0.782	0.74	0.792	0.798	0.787
GnoSSEA	0.846	0.86	0.832	0.87	0.874	0.866

Çizelge 5.5, düşük anlam karmaşıklığına sahip değerlendirme veri kümesindeki genel ölçüm sonuçlarını işaret etmektedir. Bu değerlendirme sonuçları değerlendirme veri setinin düşük belirsizliği nedeniyle açıkça daha yüksek başarım tüm sistemler tarafında elde edilmiştir. Mikro ve Makro F1 ölçümleri tez kapsamında önerilen yöntemin diğer son teknoloji sistemlere göre biraz daha üstün bir yöntem olduğunu göstermektedir. Önerilen yöntemin son derece belirsiz ortamlarda daha iyi performans gösterdiği açıkça belirtilmektedir.

5.3 Sonuç ve Değerlendirme

Bu tez kapsamında önerilen çalışma temel olarak alana özgü Varlık Bağlama görevi için sıralı bir öğrenme yöntemi sunmaktadır. Önerilen çalışmanın yapay sinir ağı makine çeviri yöntemini bir anlam ayrımı sorununa dönüştürmektedir. Ayrıca, alan bilgisinden yararlanan ve söz konusu varlık-atıf çiftlerinin kolay eşleşmelerini ortadan kaldıran aday varlıkları filtrelemektedir. Değerlendirme süreci GERBIL'deki mevcut çalışmalarla karşılaştırmak için alana özel veri setleri oluşturulması ile başlamıştır. Daha sonra GERBIL aracına uygun biçimde elde edilen değerlendirme veri kümeleri ile son teknoloji sistemler karşılaştırılmıştır. Bu karşılaştırma sonucunda bu tez kapsamında önerilen sistem alana özgü konfigürasyonda en son teknoloji yöntemlerden daha iyi performans göstermektedir.

Değerlendirme aşaması alana özgü veri kümelerinin oluşturulması ve son teknoloji Varlık Bağlama sistemlerinin internet üzerinde açık olanlarını karşılaştırabilmek için geliştirilen GERBIL aracında sınanması olmak üzere iki ana başlık altında toplanmıştır. Atıf etiketli metinlerin Vikipedi kaynağından

seçilen sinema, kitap ve müzik alanlarına göre çekilmesinden sonra bu atıflara eşleşen DBpedia varlıklarını içeren aynı alana ait bilgi tabanları çıkarılmıştır. Anlam karmaşıklığını ayarlayarak daha gerçek zamanlı ortam sağlanmıştır. Anlam karmaşıklığı çok yüksek olan sinema alanına özelleşmiş değerlendirme kümesine ek olarak anlam karmaşıklığı daha az yoğun ve çok düşük olan iki farklı veri kümesi sinema, kitap ve müzik alanlarının harmanlanması ile elde edilmiştir.

Gelecekte, gerçek zamanlı dizi oluşturma yöntemi (Akten and Grierson, 2016) yöntemini kullanarak NIL varlık problemindeki varlıkları netleştirmek için önerilen çalışma genişletilecektir. Ayrıca, genişletilmiş gerçek zamanlı yöntemi Türkçe metinlere uygulanacaktır. Daha sonra önerilen çalışma Wikipedia ve DBpedia tarafından desteklenen diğer diller için internet üzerinde erişime açılacaktır. Bu gelecek planlarını uygulama adına mevcut tez kapsamında elde edilmiş veri kümesi geliştirme aşamaları çok dil desteği sağlanarak genişletilecektir.

6 SONUÇLAR

Bu bölümde tezin kısa bir özetine ve literatüre yapılan katkılara değinilerek, tez çalışmasında elde edilen sonuçlar tartışılacak, açık problemler ve ileriki çalışmalar ele alınacaktır.

6.1 Özet ve Katkılar

Bu tezde alan uyumlu Varlık Bağlama problemi için geliştirilen diziden diziye öğrenme modeline alan bilgisini harmanlayan yeni bir ölçeklenebilir mimari önerisi getirilmiştir. Önerilen mimari, diziden diziye öğrenme modellerinden iki yönlü UKHA ve dikkat mekanizmasını kullanarak PyTorch kütüphanesinde gerçekleştirilmiştir. Gerçekleştirilen bu modeller için alan uyumlu anlamsal ve doküman gömme modülleri oluşturularak girdi olarak kullanılmıştır. Bu modülleri oluşturmak adına önceden eşlenmiş atıf-varlık çiftlerini barındıran atıf etiketli belge yığını olarak Vikipedi kaynağı kullanılırken bu atıflara eşleşmiş varlıkları içeren DBpedia bilgi tabanı sinema, kitap ve müzik olmak üzere seçilen üç alan için ayrı ayrı çıkarılmıştır.

Doküman ve anlamsal gömme modülleri sırasıyla alan tespiti ve global varlık çözümleme aşamalarında kullanılmıştır. Alan tespiti için kullanılan yöntem Doc2Vec modeli olmuştur. Birçok alan tespiti yöntemi olmasına rağmen güncel olan Doc2Vec modeli diğer yöntemlere karşı başarı sağladığı için bu tez kapsamında bu model üzerinde durulmuştur. Yapısal olmayan metin bu aşamadan geçtiğinde bağlam olarak hangi alanda olduğu tespit edilmesi ile birlikte bir sonraki aşama olan global varlık çözümleme aşamasına geçilmektedir. Global varlık çözümleme aşamasında alan tespiti yapıldığı için aday varlıkların sayısı azalmış olmaktadır. Bu durum da global varlık çözümleme aşamasını belirli bir alan için yapılmasından dolayı işlem süresini azaltmaktadır.

Son aşamada tez kapsamında önerilen yöntem güncel ve veb servisi olan Varlık Bağlama sistemleri ile GERBIL aracında karşılaştırılmıştır. Karşılaştırmanın alan uyumlu veri kümelerinde yapılabilmesi için geliştirilen WeDGeM aracı ile seçilen üç alanı da kapsayan anlam karmaşıklığı düşük ve yüksek olmak üzere iki farklı değerlendirme kümesi hazırlanmıştır. Vikipedi kaynağındaki anlamsal kategori sayfalarındaki bilgilerle DBpedia kaynağında yer alan şema bilgileri örtüştürerek alan bilgilerine göre iki kaynak için ayrı ayrı alan uyumlu veri kaynakları üretilmiştir. Sonuçlara bakıldığında bu tez kapsamında önerilen bütünüyle bakıldığında özgün yöntemin diğer bilinen yöntemlerden başarılı olduğu

gözlenmiştir.

6.2 Kısıtlamalar

Tez kapsamında önerilen ölçeklenebilir mimarinin gerçekleştiriminin genişletilebileceği pek çok yön bulunmaktadır. Öncelikle alan tespitinde belge gömme modülü için seçilen Doc2Vec modelinin güncel çalışmalarla kıyaslanarak başarısı gözlenebilir. Bu bağlamda geleneksel yöntemler ve güncel yöntemler harmanlanarak daha iyi sonuç verebilecek karma modelin geliştirilebilinir. Mevcut durumda Doc2Vec modelinin tez kapsamında seçilmesinin nedeni alan tespiti için diğer çalışmalara üstünlüğünün gösterilmesidir.

Sinir ağı modellerinin girdilerinden ikincisi olan bilgi tabanı gömme modülü için tez kapsamında RDF2Vec modeli seçilmiştir. Ancak çizge gömme modelleri çok fazla çalışılan bir alan olduğu için diğer son teknoloji çizge gömme modelleri ile tez kapsamında seçilen RDF2Vec modelinin karşılaştırılması yapılabilir. Tezin içeriğinde çoğunlukla RDF elemanlarından yararlanıldığı için bu aşamada öncelikle RDF2Vec modeli tercih edilmiştir.

Anlamsal gömme modülünde kullanılan RDF gömme modelinde tez kapsamında sadece RDF elemanlarından atıf ve varlık çiftleri özne ve nesne durumları açısından dizilişleri incelenmiştir. Atıf ve varlıklar arasında bilgi tabanındaki ilişkiler mevcut tez kapsamında değinilmemiştir. İlişki yapısının genel çözümleme yöntemindeki etkisi gözlenebilir.

Önerilen mimarinin en önemli özelliği farklı pek çok bileşeni üzerinden genişletilebilir yapıda olmasıdır. Dolayısıyla, tez çalışmasının gerçekleştiriminin devam edebileceği pek çok alan bulunmaktadır. Farklı doküman yığınları ile farklı bilgi tabanları kullanılarak tez kapsamı dışındaki alanlar için doküman ve bilgi gömme modelleri geliştirilebilir. Bu aşamada doküman yığınındaki atıflar ile bilgi tabanındaki bu atıflara karşılığı gelen varlıkların öncelikle eşlenmesi gerekebilir. Bu tez kapsamında bu problemin ana konunun dışında olmasından dolayı Vikipedi ve DBpedia kaynaklarının zaten birbiri ile atıf-varlık çiftleri açısından bağlı olduğu için bu iki kaynak kullanılmıştır.

Değerlendirme veri kümesi oluşturan araç olan WeDGeM aracının sonradan adapte edilen alan doğrultusunda genişletilerek o alana özgü de değerlendirme kümesi oluşturulabilir. Önerilen yöntemin diğer en son teknolojik sistemlerle karşılaştırabilmesi için değerlendirme kümesinin Gerbil aracına uygun veriler

biçiminde üretilmesi gerekmektedir.

Son olarak tez kapsamında İngilizce doğal diline ait metinlerde atıf ve varlık çiftlerinin bağlanması ile uğraşılmıştır. Çok dil desteğinin öncelikle Türkçe gibi diğer doğal diller için verilebilir. Anlamsal ve belge gömme modellerinin elle oluşturulmuş dil kurallarından bağımsız olması avantaj sağlamaktadır.

6.3 İleriki Çalışmalar

Bu tez kapsamında diziden diziye öğrenme metodlarına dayanan alana özgü bir Varlık Bağlama sistemi geliştirilmiştir. Sinir ağı makine çeviri çalışmasının mantığı varlık çözümleme problemine dönüştürülmüştür. Aynı zamanda alan bilgisi kullanılarak aday varlıkların sayısı filtrelenmiş ve bir tane aday varlığı olan kolay atıf-varlık eşleşmeleri de çözümleme aşamasında bağlamı tespit etme aşamasında kullanılmıştır. GERBIL aracındaki var olan Varlık Bağlama sistemleri ile üretilen alana özgü değerlendirme kümelerinde bu tezde sunulan metod karşılaştırılmıştır. Bu tez kapsamında sunulan metod internet üzerinde erişime açık olan diğer Varlık Bağlama sistemlerinde alana özgü ortamda daha başarılı sonuçlar vermiştir.

Gelecek çalışmalarda Türkçe metinlerde çalışacak şekilde sunulan yöntem genişletilecektir. Bunun için Türkçe dilindeki Vikipedi yığınları ilk aşamada kullanılacaktır. Ayrıca DBpedia bilgi tabanı Türkçe dili için farklı veri kaynaklarından zenginleştirilecektir. Bilgisayar kümelerindeki veri modelleri İngilizce dili için öncelikle gerçekleştirilecektir. Sonraki aşamada Türkçe dilinin sisteme adapte edilmesi sağlanacaktır. Vikipedi ve DBpedia kaynaklarının sağladığı herhangi bir alanda ve herhangi bir dilde Varlık Bağlama sisteminin geliştirilmesine olanak sağlanacaktır. Ayrıca farklı kaynaklardan uygun verilerin mevcut yöntemle etiketlenerek Vikipedi ve DBpedia kaynaklarından bağımsız bir altyapının kurulması sağlanacaktır. Bu şekilde Türkçe metinlerden oluşan alana özgü doküman ve anlamsal gömme modelleri internet üzerinden erişime açılarak metin üzerine yapılacak diğer yöntemler için de öncü bir çalışma olacaktır.

Varlık-İlişki gömme modelleri çıkarılırken farklı çizge çekirdeklerinden faydanılmaktadır. Ayrıca bu aşamada çizge çekirdekleri incelenerek mevcut başarımı artırabilecek yöntemler araştırılacaktır. Rastgele çizge üzerinde dolaşmak yerine belirli bir çekirdek kümeden yola çıkarak seçilen alan için gerekli en iyi ontoloji yapısını oluşturmak hedeflenecektir.

Son olarak büyük veri ortamının kurulması ve bilgisayar kümelerine

aktarılması planlanmaktadır. Bu durumda RDF-HDT (Martínez-Prieto et al., 2012) aracı DBpedia ve LinkedMDB üzerinde çalışabilir hale getirilecektir. Böylece sorgulanabilir hızlı ve etkin bir yerel uç noktayı bölüm bünyesine dahil edilecek ve ileriki çalışmalar için zemin hazırlanacaktır. Buna ek olarak Vikipedi yığınlarının Akka aracı sayesinde dağıtık bir şekilde doküman veri deposuna aktarımı gerçekleştirilecektir. Ek olarak DBpedia etiket verisi de anahtar-değer veri deposuna aktarılacaktır. Tüm bu işlemler docker[42] aracı ile esnek bir şekilde yürütülecektir. Bilgisayar kümelerinden oluşan bu veri ortamı sağlandıktan sonra ontoloji yapısındaki ve içeriğe dayalı gömme modellerini melezleyen özgün Varlık Bağlama yönteminin geliştirilmesi aşamasında mevcut tez kapsamında en yüksek performansı gösteren ontoloji yapısı seçilecektir.

[42]https://www.docker.com/

KAYNAKLAR DİZİNİ

Agirre, E. and Soroa, A.: 2009, Personalizing PageRank for word sense disambiguation, in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics, 33–41 pp

Akten, M. and Grierson, M.: 2016, Real-time interactive sequence generation and control with recurrent neural network ensembles, *arXiv preprint arXiv:1612.04687*

Alpaydin, E.: 2014, *Introduction to machine learning*, MIT press

Arcan, M., Turchi, M., Tonelli, S., and Buitelaar, P.: 2014, Enhancing statistical machine translation with bilingual terminology in a cat environment, in *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, 54–68 pp

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z.: 2007, Dbpedia: A nucleus for a web of open data, in *The semantic web*, Springer, 722–735 pp

Bahdanau, D., Cho, K., and Bengio, Y.: 2014, Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473*

Bengio, Y., Simard, P., and Frasconi, P.: 1994, Learning long-term dependencies with gradient descent is difficult, *IEEE transactions on neural networks* **5(2)**, 157

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S.: 2009, Dbpedia-a crystallization point for the web of data, *Web Semantics: science, services and agents on the world wide web* **7(3)**, 154

Blei, D. M., Ng, A. Y., and Jordan, M. I.: 2003, Latent dirichlet allocation, *Journal of machine Learning research* **3(Jan)**, 993

Bodenreider, O.: 2004, The unified medical language system (umls): integrating biomedical terminology, *Nucleic acids research* **32(suppl_1)**, D267

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J.: 2008, Freebase: a collaboratively created graph database for structuring human knowledge, in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, AcM, 1247–1250 pp

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O.: 2013, Translating embeddings for modeling multi-relational data, in *Advances in neural information processing systems*, 2787–2795 pp

KAYNAKLAR DİZİNİ (devam)

Bunescu, R. and Paşca, M.: 2006, Using encyclopedic knowledge for named entity disambiguation, in *11th conference of the European Chapter of the Association for Computational Linguistics*

Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., and Trani, S.: 2013, Dexter: an open source framework for entity linking, in *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, ACM, 17–20 pp

Chang, K.-W., Yih, S. W.-t., Yang, B., and Meek, C.: 2014, Typed tensor decomposition of knowledge bases for relation extraction

Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., and Johnson, M.: 2000, Bllip 1987-89 wsj corpus release 1, *Linguistic Data Consortium, Philadelphia* 36

Chen, Z., Tamang, S., Lee, A., Li, X., Lin, W.-P., Snover, M. G., Artiles, J., Passantino, M., and Ji, H.: 2010, Cuny-blender tac-kbp2010 entity linking and slot filling system description., in *TAC*

Cochez, M., Ristoski, P., Ponzetto, S. P., and Paulheim, H.: 2017, Biased graph walks for rdf graph embeddings, in *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, ACM, 21 p.

Cornolti, M., Ferragina, P., and Ciaramita, M.: 2013, A framework for benchmarking entity-annotation systems, in *Proceedings of the 22nd international conference on World Wide Web*, ACM, 249–260 pp

Cucerzan, S.: 2007a, Large-scale named entity disambiguation based on wikipedia data, in *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, 708–716 pp

Cucerzan, S.: 2007b, Large-scale named entity disambiguation based on wikipedia data

Dai, A. M., Olah, C., and Le, Q. V.: 2015, Document embedding with paragraph vectors, *arXiv preprint arXiv:1507.07998*

Dalvi, N., Kumar, R., and Pang, B.: 2012, Object matching in tweets with spatial models, in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, New York, NY, USA, ACM, 43–52 pp

de Melo, G. and Weikum, G.: 2012, Uwn: A large multilingual lexical knowledge base, in *Proceedings of the ACL 2012 System Demonstrations*, Association for

KAYNAKLAR DİZİNİ (devam)

Computational Linguistics, 151–156 pp

de Vries, G. K.: 2013, A fast approximation of the weisfeiler-lehman graph kernel for rdf data, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 606–621 pp

de Vries, G. K. D. and de Rooij, S.: 2015, Substructure counting graph kernels for machine learning from rdf data, *Web Semantics: Science, Services and Agents on the World Wide Web* **35(Part 2)**, 71 , Machine Learning and Data Mining for the Semantic Web (MLDMSW)

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W.: 2014, Knowledge vault: A web-scale approach to probabilistic knowledge fusion, in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 601–610 pp

Dredze, M., McNamee, P., Rao, D., Gerber, A., and Finin, T.: 2010, Entity disambiguation for knowledge base population, in *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 277–285 pp

D'Souza, J. and Ng, V.: 2015, Sieve-based entity linking for the biomedical domain, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2, 297–302 pp

Ellis, J., Getman, J., Mott, J., Li, X., Griffitt, K., Strassel, S., and Wright, J.: 2013, Linguistic resources for 2013 knowledge base population evaluations, in *Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013*

Ernst, P., Siu, A., and Weikum, G.: 2015, Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences, *BMC Bioinformatics* **16**, 157:1

Fang, W., Zhang, J., Wang, D., Chen, Z., and Li, M.: 2016, Entity disambiguation by knowledge and text jointly embedding, in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 260–269 pp

Ferragina, P. and Scaiella, U.: 2010, Tagme: On-the-fly annotation of short text fragments (by wikipedia entities), in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, New York, NY, USA, ACM, 1625–1628 pp

KAYNAKLAR DİZİNİ (devam)

Ferragina, P. and Scaiella, U.: 2012, Fast and accurate annotation of short texts with wikipedia pages, *IEEE software* **29(1)**, 70

Francis, W. and Kucera, H.: 1982, Frequency analysis of english usage

Gale, W. A., Church, K. W., and Yarowsky, D.: 1992, A method for disambiguating word senses in a large corpus, *Computers and the Humanities* **26(5)**, 415

Ganea, O.-E., Ganea, M., Lucchi, A., Eickhoff, C., and Hofmann, T.: 2016, Probabilistic bag-of-hyperlinks model for entity linking, in *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 927–938 pp

Graves, A. and Schmidhuber, J.: 2005, Framewise phoneme classification with bidirectional lstm and other neural network architectures, *Neural Networks* **18(5-6)**, 602

Gupta, N., Singh, S., and Roth, D.: 2017, Entity linking via joint encoding of types, descriptions, and context, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2681–2690 pp

Han, X. and Sun, L.: 2011, A generative entity-mention model for linking entities with knowledge base, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, 945–954 pp

Han, X., Sun, L., and Zhao, J.: 2011, Collective entity linking in web text: a graph-based method, in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, ACM, 765–774 pp

Hassanzadeh, O. and Consens, M. P.: 2009, Linked movie data base., in *LDOW*

He, Z., Liu, S., Li, M., Zhou, M., Zhang, L., and Wang, H.: 2013, Learning entity representation for entity disambiguation, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, 30–34 pp

Hochreiter, S. and Schmidhuber, J.: 1997, Long short-term memory, *Neural computation* **9(8)**, 1735

Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., De Melo, G., and Weikum, G.: 2011a, Yago2: exploring and querying world knowledge in time, space, context, and many languages, in *Proceedings of the 20th international conference companion on World wide web*, ACM, 229–232 pp

KAYNAKLAR DİZİNİ (devam)

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G.: 2011b, Robust disambiguation of named entities in text, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 782–792 pp

Huang, J., Zhou, M., and Yang, D.: 2007, Extracting chatbot knowledge from online discussion forums., in *IJCAI*, Vol. 7, 423–428 pp

Inan, E. and Dikenelli, O.: 2017, Wedgem: A domain-specific evaluation dataset generator for multilingual entity linking systems, in *International Conference on Web Information Systems Engineering*, Springer, 221–228 pp

Ji, G., He, S., Xu, L., Liu, K., and Zhao, J.: 2015, Knowledge graph embedding via dynamic mapping matrix, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, 687–696 pp

Kulkarni, S., Singh, A., Ramakrishnan, G., and Chakrabarti, S.: 2009, Collective annotation of wikipedia entities in web text, in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 457–466 pp

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K.: 2015, From word embeddings to document distances, in *International Conference on Machine Learning*, 957–966 pp

Lafferty, J., McCallum, A., and Pereira, F. C.: 2001, Conditional random fields: Probabilistic models for segmenting and labeling sequence data

Le, Q. and Mikolov, T.: 2014, Distributed representations of sentences and documents, in *International Conference on Machine Learning*, 1188–1196 pp

Ley, M.: 2002, The dblp computer science bibliography: Evolution, research issues, perspectives, in *International symposium on string processing and information retrieval*, Springer, 1–10 pp

Li, X., Strassel, S. M., Ji, H., Griffitt, K., and Ellis, J.: 2012, Linguistic resources for entity linking evaluation: from monolingual to cross-lingual, *Annotation* **1**, 1

Li, Y., Wang, C., Han, F., Han, J., Roth, D., and Yan, X.: 2013, Mining evidences for named entity disambiguation, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*,

KAYNAKLAR DİZİNİ (devam)

ACM, 1070–1078 pp

Lipczak, M., Koushkestani, A., and Milios, E.: 2014, Tulip: Lightweight entity recognition and disambiguation using wikipedia-based topic centroids, in *Proceedings of the first international workshop on Entity recognition & disambiguation*, ACM, 31–36 pp

Liu, Y., Shen, W., and Yuan, X.: 2016, Deola: a system for linking author entities in web document with dblp, in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ACM, 2449–2452 pp

Mahdisoltani, F., Biega, J., and Suchanek, F. M.: 2013, Yago3: A knowledge base from multilingual wikipedias, in *CIDR*

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D.: 2014, The stanford corenlp natural language processing toolkit, in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60 pp

Martínez-Prieto, M. A., Gallego, M. A., and Fernández, J. D.: 2012, Exchange and consumption of huge rdf data, in *Extended Semantic Web Conference*, Springer, 437–452 pp

Melville, P., Gryc, W., and Lawrence, R. D.: 2009, Sentiment analysis of blogs by combining lexical knowledge with text classification, in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 1275–1284 pp

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C.: 2011, Dbpedia spotlight: Shedding light on the web of documents, in *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, New York, NY, USA, ACM, 1–8 pp

Mikolov, T., Chen, K., Corrado, G., and Dean, J.: 2013a, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*

Mikolov, T., Chen, K., Corrado, G., and Dean, J.: 2013b, Efficient estimation of word representations in vector space, *CoRR* abs/1301.3781

Miller, G. A.: 1995, Wordnet: a lexical database for english, *Communications of the ACM* **38(11)**, 39

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T.: 1993, A semantic concordance, in *Proceedings of the workshop on Human Language Technology*, Association for Computational Linguistics, 303–308 pp

KAYNAKLAR DİZİNİ (devam)

Milne, D. and Witten, I. H.: 2008, Learning to link with wikipedia, in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, New York, NY, USA, ACM, 509–518 pp

Minervini, P., d'Amato, C., Fanizzi, N., and Esposito, F.: 2016, Leveraging the schema in latent factor models for knowledge graph completion, in *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016*, 327–332 pp

Mitchell, A., Strassel, S., Huang, S., and Zakhary, R.: 2005, Ace 2004 multilingual training corpus, *Linguistic Data Consortium, Philadelphia* **1**, 1

Moro, A., Cecconi, F., and Navigli, R.: 2014a, Multilingual word sense disambiguation and entity linking for everybody, in *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, Aachen, Germany, Germany, CEUR-WS.org, 25–28 pp

Moro, A., Raganato, A., and Navigli, R.: 2014b, Entity Linking meets Word Sense Disambiguation: a Unified Approach, *Transactions of the Association for Computational Linguistics (TACL)* **2**, 231

Navigli, R.: 2009, Word sense disambiguation: A survey, *ACM computing surveys (CSUR)* **41(2)**, 10

Navigli, R.: 2013, Babelnet and friends: A manifesto for multilingual semantic processing, *Intelligenza Artificiale* **7(2)**, 165

Navigli, R. and Ponzetto, S. P.: 2012a, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial Intelligence* **193**, 217

Navigli, R. and Ponzetto, S. P.: 2012b, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial Intelligence* **193**, 217

Nguyen, D. B., Hoffart, J., Theobald, M., and Weikum, G.: 2014a, Aida-light: High-throughput named-entity disambiguation., in C. Bizer, T. Heath, S. Auer, and T. Berners-Lee (eds.), *LDOW*, Vol. 1184 of *CEUR Workshop Proceedings*, CEUR-WS.org

Nguyen, D. B., Hoffart, J., Theobald, M., and Weikum, G.: 2014b, Aida-light: High-throughput named-entity disambiguation., in *LDOW*

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E.: 2015, A review of relational machine learning for knowledge graphs: From multi-relational link prediction to automated knowledge graph construction, *CoRR* abs/1503.00759

KAYNAKLAR DİZİNİ (devam)

Noia, T. D., Ostuni, V. C., Rosati, J., Tomeo, P., Sciascio, E. D., Mirizzi, R., and Bartolini, C.: 2016, Building a relatedness graph from linked open data: A case study in the IT domain, *Expert Syst. Appl.* **44**, 354

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B.: 2010, Semantic similarity and relatedness between clinical terms: an experimental study, in *AMIA annual symposium proceedings*, Vol. 2010, American Medical Informatics Association, 572 p.

Pamay, T., Sulubacak, U., Torunoglu-Selamet, D., and Eryigit, G.: 2015, The annotation process of the itu web treebank, in *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, 95 p.

Pantel, P. and Fuxman, A.: 2011, Jigs and lures: Associating web queries with structured entities, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics, 83–92 pp

Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G.: 2007, Measures of semantic similarity and relatedness in the biomedical domain, *Journal of biomedical informatics* **40(3)**, 288

Phan, M. C., Sun, A., Tay, Y., Han, J., and Li, C.: 2017, Neupl: Attention-based semantic matching and pair-linking for entity disambiguation, in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, 1667–1676 pp

Piccinno, F. and Ferragina, P.: 2014, From tagme to WAT: a new entity annotator, in *ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia*, 55–62 pp

Ratinov, L., Roth, D., Downey, D., and Anderson, M.: 2011, Local and global algorithms for disambiguation to wikipedia, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, 1375–1384 pp

Ristoski, P. and Paulheim, H.: 2016, Rdf2vec: RDF graph embeddings for data mining, in *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*, 498–514 pp

Sak, H., Güngör, T., and Saraçlar, M.: 2008, Turkish language resources: Morphological parser, morphological disambiguator and web corpus, in *GoTAL*

KAYNAKLAR DİZİNİ (devam)

2008, Vol. 5221 of *LNCS*, Springer, 417–427 pp

Salton, G. and Buckley, C.: 1988, Term-weighting approaches in automatic text retrieval, *Information processing & management* **24(5)**, 513

Schuhmacher, M. and Ponzetto, S. P.: 2014, Knowledge-based graph document modeling, in *Proceedings of the 7th ACM international conference on Web search and data mining*, ACM, 543–552 pp

Shadbolt, N., Berners-Lee, T., and Hall, W.: 2006, The semantic web revisited, *IEEE intelligent systems* **21(3)**, 96

Shen, W., Han, J., Wang, J., Yuan, X., and Yang, Z.: 2018, Shine+: A general framework for domain-specific entity linking with heterogeneous information networks, *IEEE Transactions on Knowledge and Data Engineering* **30(2)**, 353

Shen, W., Wang, J., and Han, J.: 2015, Entity linking with a knowledge base: Issues, techniques, and solutions, *Transactions on Knowledge and Data Engineering* **27(2)**, 443

Shen, W., Wang, J., Luo, P., and Wang, M.: 2012, Linden: linking named entities with knowledge base via semantic knowledge, in *Proceedings of the 21st international conference on World Wide Web*, ACM, 449–458 pp

Singh, S., Subramanya, A., Pereira, F., and McCallum, A.: 2012, Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia, *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012-015*

Spitkovsky, V. I. and Chang, A. X.: 2012, A cross-lingual dictionary for english wikipedia concepts., in *LREC*, 3168–3175 pp

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D.: 2006, The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages, *arXiv preprint cs/0609058*

Strassel, S., Przybocki, M. A., Peterson, K., Song, Z., and Maeda, K.: 2008, Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction., in *LREC*

Suchanek, F. M., Kasneci, G., and Weikum, G.: 2007, Yago: a core of semantic knowledge, in *Proceedings of the 16th international conference on World Wide Web*, ACM, 697–706 pp

Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., and Wang, X.: 2015, Modeling mention, context and entity with neural networks for entity disambiguation., in *IJCAI*, 1333–1339 pp

KAYNAKLAR DİZİNİ (devam)

Sutskever, I., Tenenbaum, J. B., and Salakhutdinov, R. R.: 2009, Modelling relational data using bayesian clustered tensor factorization, in *Advances in neural information processing systems*, 1821–1828 pp

Sutskever, I., Vinyals, O., and Le, Q. V.: 2014, Sequence to sequence learning with neural networks, in *Advances in neural information processing systems*, 3104–3112 pp

Tjong Kim Sang, E. F. and De Meulder, F.: 2003, Introduction to the conll-2003 shared task: Language-independent named entity recognition, in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Association for Computational Linguistics, 142–147 pp

Usbeck, R., Ngomo, A.-C. N., Röder, M., Gerber, D., Coelho, S., Auer, S., and Both, A.: 2014, AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data, in P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble (eds.), *The Semantic Web – ISWC 2014*, Vol. 8796 of *Lecture Notes in Computer Science*, Springer International Publishing, 457-471 pp

Usbeck, R., Röder, M., Ngonga Ngomo, A.-C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., and Wesemann, L.: 2015, GERBIL – general entity annotation benchmark framework, in *24th WWW conference*

Waitelonis, J. and Sack, H.: 2016, Named entity linking in# tweets with kea., in *# Microposts*, 61–63 pp

West, R., Gabrilovich, E., Murphy, K., Sun, S., Gupta, R., and Lin, D.: 2014, Knowledge base completion via search-based question answering, in *Proceedings of the 23rd international conference on World wide web*, ACM, 515–526 pp

Weston, J., Bordes, A., Yakhnenko, O., and Usunier, N.: 2013, Connecting language and knowledge bases with embedding models for relation extraction, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, 1366–1371 pp

Yamada, I., Shindo, H., Takeda, H., and Takefuji, Y.: 2016, Joint learning of the embedding of words and entities for named entity disambiguation, *arXiv preprint arXiv:1601.01343*

KAYNAKLAR DİZİNİ (devam)

Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L.: 2014, Learning multi-relational semantics using neural-embedding models, *arXiv preprint arXiv:1411.4072*

Zhang, J., Li, J., Li, X.-L., Shi, Y., Li, J., and Wang, Z.: 2016, Domain-specific entity linking via fake named entity detection, in *International Conference on Database Systems for Advanced Applications*, Springer, 101–116 pp

Zhong, Z. and Ng, H. T.: 2010, It makes sense: A wide-coverage word sense disambiguation system for free text, in *Proceedings of the ACL 2010 System Demonstrations*, Association for Computational Linguistics, 78–83 pp

Zwicklbauer, S., Seifert, C., and Granitzer, M.: 2016a, *DoSeR - A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings*, 182–198 pp, Springer International Publishing, Cham

Zwicklbauer, S., Seifert, C., and Granitzer, M.: 2016b, Robust and collective entity disambiguation through semantic embeddings, in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, ACM, 425–434 pp

ÖZGEÇMİŞ

Adı Soyadı: Emrah İNAN
Doğum Tarihi: 1985
Doğum Yeri: Ankara
Uyruğu: T.C.

EĞİTİM

Yüksek Lisans: İzmir Yüksek Teknoloji Enstitüsü
Bilgisayar Mühendisliği Bölümü, 2012

Lisans: İzmir Ekonomi Üniversitesi
Yazılım Mühendisliği Bölümü, 2008

Lise: Fen Lisesi, Çorum, 2003

İLGİ ALANLARI

Doğal Dil İşleme, Bağlı Açık Veri, Makine Öğrenme, Yapay Zeka